BIQ: Bisection Interval Quantization for Communication-efficient Federated Learning

Luyang Gai[1, 2], **Shusen Yang**[1, 2*], **Xuebin Ren**[1, 3], **Zihao Zhou**[1, 2]

[1]National Engineering Laboratory for Big Data Analytics, Xi'an Jiaotong University, China
[2]School of Mathematics and Statistics, Xi'an Jiaotong University, China
[3]School of Computer Science and Technology , Xi'an Jiaotong University, China
wnt0954@gmail.com, shusenyang@mail.xjtu.edu.cn, xuebinren@mail.xjtu.edu.cn, zihaozhou19@gmail.com

Abstract

Quantization is a pivotal technique for enhancing communication efficiency in Federated Learning (FL). Traditional quantization methods often set uniform intervals, may fail to adequately characterize non-uniform data distributions, thus leading to substantial estimation errors and degrated model performance. Non-uniform quantization can better solve the problem. However, when applied to FL, it would bring additional communication overheads for the alignment of parameter distributions among distributed models. To address this issue, we propose Bisection Interval Quantization (BIQ), a novel non-uniform quantization framework for FL with great communication efficiency. In particular, BIQ works by optimizing the interval selection through recursive bisection among distributed clients without extra parameter communication. For scenarios involving amounts of boundary inputs, we further design Weighted Bisection Interval Quantization (WBIQ), which incorporates maximum likelihood estimation to refine boundary value reconstruction to enhance the estimation quality of boundary inputs. Our theoretical analysis rigorously establishes, for the first time under biased quantization conditions, that both BIQ and WBIQ achieve tighter error bounds and enhanced stability. Extensive experiments validate that both BIQ and WBIQ significantly accelerate the convergence of FL training when compared to the state-of-the-art quantizers under both convex and non-convex settings.

Code and Extended version —
https://github.com/GaiLuyang/BIQ

Introduction

Federated Learning (FL) (McMahan et al. 2017) has emerged as a promising paradigm for distributed machine learning, particularly in privacy-sensitive domains such as healthcare (Pati et al. 2022; Boscarino et al. 2022) and finance (Cui et al. 2021). FL enables collaborative model training across distributed clients while preserving data privacy. However, its practical deployment is hindered by communication bottlenecks caused by frequent parameter exchanges. To address it, quantization (Elgabli et al. 2020; Reisizadeh et al. 2020; Sun et al. 2022) has become a cornerstone for reducing communication overhead, and most algorithms predominantly rely on uniform quantizers including Stochastic Quantization (SQ) (Elgabli et al. 2020; Reisizadeh et al. 2020; Gupta et al. 2015; Alistarh et al. 2017) and Rounding Quantization (RQ) (Sun et al. 2022; Gupta et al. 2015; Bai, Wang, and Liberty 2018; Nagel et al. 2022). However, these methods enforce equal intervals that fail to adapt to most neural networks with non-uniform parameter distributions (Han, Mao, and Dally 2015; Zhang et al. 2018; Jung et al. 2019) (e.g., such as bell-shaped and long-tail distributions observed in weights and activations (Gongyo et al. 2024; Li, Dong, and Wang 2019)), resulting in substantial information loss and degraded convergence.

Non-uniform quantization (Zhang et al. 2018; Chen et al. 2023; Luqman, Qazi, and Khan 2024; Wang et al. 2022) offers superior accuracy compared to uniform methods by dynamically mapping the parameters to non-uniformly spaced quantized values to match non-uniform distributions. Prior works, such as loss-driven adaptive quantization (Jung et al. 2019), power-of-two quantization (Li, Dong, and Wang 2019), logarithmic quantization (Miyashita, Lee, and Murmann 2016) and iterative range-level updates (Luqman, Qazi, and Khan 2024), have demonstrated its effectiveness in centralized settings. However, non-uniform quantization poses an inherent deployment challenge to FL due to the nature of misalignment of adaptive quantization schemes across heterogeneous clients. One solution is that clients upload both quantized parameters and their respective quantization schemes (Chen et al. 2023). However, this approach introduces extra communication and computational costs, thus undermining the communication benefits of quantization. What's more, additional parameters may expose information such as data distribution of clients, resulting in privacy leakage. Therefore, there is a pressing need to design non-uniform quantization schemes for FL to improve model accuracy without incurring additional communication costs.

The key insight that inspires our work lies in the object of quantization. Quantization in FL has been largely inherited from model quantization, where the quantized outputs must retain semantic meaning because they are directly used in training and inference. However, in the context of FL, quantization serves solely as a communication compression mechanism, as illustrated in Figure 1. This inspires us to design novel encoding–decoding mechanisms tailored specifically for the communication process. As a result, quanti-

*Corresponding author.

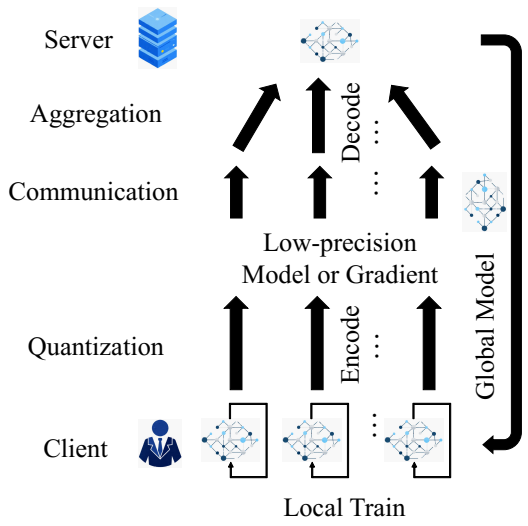

Figure 1: Quantification in FL. Quantization applies to client encoding and server decoding processes.

zation in FL opens up a broader design space compared to traditional model quantization. We revisit the fundamental interpretation of quantization levels: in most prior works, quantization levels serve as symbolic representations of data points. This naturally raises a question: Can we assign quantization a new meaning?

To address the problem, we try to quantize the process of quantization. Building on this, we propose Bisection Interval Quantization (BIQ), a novel non-uniform quantizer that recursively bisects intervals using binary codes. Each bit in BIQ encodes a client's local bisection path, eliminating the need for explicit interval alignment. For boundary-critical scenarios (e.g., extreme gradient values), we further design Weighted BIQ (WBIQ), which refines boundary reconstruction through maximum likelihood estimation based on bit-count statistics. Crucially, both methods operate without increasing communication overhead.

However, in terms of theoretical analysis, BIQ and WBIQ sacrifice unbiasedness for higher quantization accuracy. Theoretical guarantees for biased quantizers in FL have remained elusive, as prior convergence analyses (Elgabli et al. 2020; Reisizadeh et al. 2020) strictly require unbiased quantization. We bridge this gap by establishing the first convergence bounds for biased quantizers under both strongly convex and non-convex objectives.

Our Contribution. In summary, our contributions are as follows: First, we propose a novel BIQ algorithm for FL, which achieves higher quantization accuracy without sacrificing communication cost. Second, we propose its variant WBIQ to further improve quantization accuracy. WBIQ incorporates maximum likelihood estimation to better capture input values near the boundaries of the quantization range. Third, we establish the first convergence analysis for biased quantizers in FL. Under appropriate control of bias and noise, we demonstrate that BIQ and WBIQ converge at rates of $\mathcal{O}\left(\frac{1}{T}\right)$ under strong convexity assumptions and $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ in non-convex settings, respectively. Fourth, extensive experiments on four datasets validate that BIQ and WBIQ achieve more accurate quantization, faster convergence and higher model accuracy compared with existing quantization methods under the same resolution.

Our work redefines the role of quantization and establishes a general analytical paradigm that accommodates diverse quantization strategies in FL. Our results unlock the potential of high-precision FL in bandwidth-limited scenarios.

Problem Definition

We first introduce FL and present the quantization-based communication compression method.

Federated Learning (FL). In FL, a server collaborates with clients $\mathcal{N} = \{1, 2, ..., N\}$ to jointly train a model by solving the following Empirical Risk Minimization (ERM) problem:

$$\min_{\theta} f(\theta) \triangleq \min_{\theta} \sum_{n=1}^{N} \frac{M_n}{N} f_n(\theta; \mathcal{M}_n), \tag{1}$$

where $\theta \in R^d$ denotes the model parameter. $\mathcal{M}_n$ is the local dataset with M_n samples. f_n represents the loss function of the n-th client.

SGD-Based Optimizer in FL with Quantization. In FL, the server distributes the initialized model θ^0 to each client. To enhance local training efficiency, the n-th client, when participating in the k-th communication round, randomly samples a mini-batch $\hat{\mathcal{M}}_n$ from its local dataset $\mathcal{M}_n$ and updates its local model according to the following procedure

$$\theta_n^{k,t+1} = \theta_{n,t}^k - \alpha \hat{\nabla} f_n(\theta_n^{k,t}; \hat{\mathcal{M}}_n), \tag{2}$$

where $t = 0, 1, ..., \tau - 1$ denotes the local update rounds. After completing these updates, the client uploads the difference $\theta_n^{k,\tau} - \theta^k$ to the server for aggregation. The communication bottleneck occurs during the update transmission process, as limited bandwidth can lead to congestion when transmitting full updates. We quantize the uploaded data to reduce the communication overhead. Additionally, resource heterogeneity poses another challenge in FL. Some clients may not respond promptly to the server's update requests (Reisizadeh et al. 2020; Sun et al. 2022), resulting in inefficient training and amplifying the impact of limited communication bandwidth. To address these issues, at each communication round, we select a subset of clients $S_k \subset \mathcal{N}$. These clients quantize the model updates by quantizer $Q(\cdot)$ and upload the quantized models for aggregation, which efficiently alleviates the communication bottleneck and improves training efficiency. The aggregated process is as follows

$$\theta^{k+1} = \theta^k + \frac{1}{s} \sum_{n \in S_k} Q(\theta_n^{k,\tau} - \theta^k), \tag{3}$$

where $|S_k| = s$. Equation (3) directly impacts the effectiveness of model aggregation. At lower quantization resolutions, quantization significantly reduces communication

Algorithm 1: BIQ encoder process.

Input: $\mathbf{x} \in \mathbb{R}^d, b, R$
Output: B

$q_L = -R \cdot \mathbf{1}, q_R = R \cdot \mathbf{1}$
for $j = 1, 2, ..., d$ **do**
  for $i = b-1, b-2, ..., 0$ **do**
    if $x^{(j)} \leq \frac{q_L^{(j)} + q_R^{(j)}}{2}$ **then**
      $q_R^{(j)} = \frac{q_L^{(j)} + q_R^{(j)}}{2}, B_i^{(j)} = 0$
    else
      $q_L^{(j)} = \frac{q_L^{(j)} + q_R^{(j)}}{2}, B_i^{(j)} = 1$
    end if
  end for
  $B^{(j)} = B_{b-1}^{(j)} B_{b-2}^{(j)} ... B_0^{(j)}$
end for

costs at the expense of increased errors. Therefore, designing high-precision quantizers at low quantization resolutions is an effective approach to achieve a better trade-off between communication efficiency and utility.

Bisection Interval Quantization

In this section, we first propose BIQ and WBIQ. Then, we present error analysis and convergence analysis.

BIQ Encoding Process

BIQ encodes the selection process of quantized values using binary system. Specifically, given a quantization resolution b, a quantization range R and an input $x \in \mathbb{R}^d$ with each coordinate satisfying $x^{(j)} \in [-R, R]$, BIQ recursively partitions the interval $[-R, R]$ into subintervals by bisection. At each step, a bit ("0" or "1") indicates whether the input lies in the left or right subinterval, respectively. BIQ performs the following process when determining the encoded value for element $x^{(j)}$: if the input lies in the left half of the current quantization interval, the corresponding bit is set to $B_j^{(j)} = 0$; if it lies in the right half, the bit is set to $B_j^{(j)} = 1$. Let $B = (B_{b-1}^{(1)} B_{b-2}^{(1)} ... B_0^{(1)}, B_{b-1}^{(2)} B_{b-2}^{(2)} ... B_0^{(2)}, ..., B_{b-1}^{(d)} B_{b-2}^{(d)} ... B_0^{(d)})$ represent the quantized binary code generated by BIQ. Algorithm 1 describes the encoding mechanism, where $\mathbf{1} = (1, 1, ..., 1) \in \mathbb{R}^d$.

BIQ Decoding Process

The client transmits the binary code B to the server, which decodes each coordinate $B^{(j)}$ according to its encoding scheme. Specifically, the server processes the bits in $B^{(j)}$ from left to right, where a "0" indicates that the left half of the current interval is retained, and a "1" indicates that the right half is retained. The decoding process of BIQ results in an interval $[q^L_{B^{(j)}}, q^R_{B^{(j)}}]$. However, the aggregation step requires a deterministic value. To minimize quantization error, BIQ selects the midpoint of the interval as the final output: $q_{B^{(j)}} = \frac{q^L_{B^{(j)}} + q^R_{B^{(j)}}}{2}$.

Algorithm 2: FedBIQ.

Input: N clients with datasets $\{\mathcal{M}_n\}_{n=1}^N$, initialize server model parameters θ^0, communication rounds K, quantization resolution b, learning rate α.
Output: Global model θ^*

Server broadcasts θ^0 to all clients.
for $k = 0, 1, ..., K-1$ **do**
  Server broadcasts θ^k to clients S^k who are selected uniformly at random.
  for client $n \in S_k$ in parallel **do**
    for $t = 0, 1, ..., \tau - 1$ **do**
      Local update via Equation (2)
    end for
    $B_n^k \leftarrow BIQ(\theta_n^{k,\tau} - \theta^k, b, R_n^k)$
    Send B_n^k, R_n^k to Server.
  end for
  Server decodes quantized signals and aggregates via Equation (3).
end for

BIQ Variant: WBIQ. Since BIQ uses the midpoint of the interval as its output, it introduces gaps when representing inputs that are close to the boundaries of the quantization range. To address this limitation while maintaining the same communication cost, we propose an enhanced scheme called WBIQ. WBIQ aims to more accurately handle the quantization of inputs near the range boundaries without increasing the communication overhead.

Definition 1 (Counting)**.** For any $c = c_{b-1} c_{b-2} ... c_0, c_i \in \{0, 1\}$, we define the functions $\delta_j : \{0, 1\}^b \to \mathbb{Z}_{\geq 0}, j \in \{0, 1\}$. δ_j count the number of j in c. Specifically, $\bar{\delta}_0(c) = b - \sum_{i=1}^b c_i, \delta_1(c) = \sum_{i=1}^b c_i$.

Our improvement is inspired by the principle of maximum likelihood estimation. In this context, the bits "0" and "1" represent the number of times the left and right intervals were selected during the bisection process, respectively. If "1" appears more frequently than "0" in $B^{(j)}$, then according to maximum likelihood estimation, the input value is more likely to lie closer to the right end of the interval. This insight motivates our design of WBIQ, where we incorporate the number of "0"s and "1"s in $B^{(j)}$ to refine the quantized output.

In WBIQ, the quantization interval $[q^L_{B^{(j)}}, q^R_{B^{(j)}}]$ is determined in the same manner as in standard BIQ. However, WBIQ further leverages the maximum likelihood estimation principle by performing a weighted average of the interval endpoints. The weights are determined by the relative frequencies of "0" and "1". Specifically, the final output is given by $q_{B^{(j)}} = \frac{\delta_0(B^{(j)})}{b} q^L_{B^{(j)}} + \frac{\delta_1(B^{(j)})}{b} q^R_{B^{(j)}}$. This maximum likelihood-based approach improves the quantization accuracy for inputs near the boundaries of the range, thereby enhancing the overall performance of the algorithm.

Error Analysis

In this section, we analyze the quantization error introduced by BIQ and WBIQ. Let C, b denote the quantization range

and the quantization resolution respectivly. After decoding, the server reconstructs the interval $[q^L_{B^{(j)}}, q^R_{B^{(j)}}]$. The true input value x_j can lie anywhere within this interval. Assuming without loss of generality that x_j is uniformly distributed over $[q^L_{B^{(j)}}, q^R_{B^{(j)}}]$, then the quantization error is $\eta_j^{BIQ} = x_j - BIQ(x_j) \sim \mathcal{U}(-\frac{q^R_{B^{(j)}} - q^L_{B^{(j)}}}{2}, \frac{q^R_{B^{(j)}} - q^L_{B^{(j)}}}{2})$. The variance of the quantization error is $D\eta_j^{BIQ} = \frac{C^2}{12 \cdot 2^{2b}}$. In contrast, for SQ, $D\eta_j^{SQ} = \frac{C^2}{6 \cdot (2^b - 1)^2}$. Comparing the two error variances yields:

$$\frac{D\eta_j^{BIQ}}{D\eta_j^{SQ}} = \frac{6 \cdot (2^b - 1)^2}{12 \cdot 2^{2b}} \leq 0.5, \quad \forall b.$$

Under the same quantization resolution b, the variance of the quantization error for BIQ is less than half of SQ, indicating that BIQ produces more stable and reliable quantized outputs. For WBIQ, we have $D\eta_j^{WBIQ} \leq D\eta_j^{SQ}$, which still ensures greater stability in the quantized results compared to SQ. In terms of error bounds, the range of the quantization error satisfies $|\eta_j^{BIQ}| \leq \frac{C}{2^b}, |\eta_j^{WBIQ}| \leq \frac{2C}{2^b}$. While for SQ, the bound is $|\eta_j^{SQ}| \leq \frac{2C}{2^b - 1}$. These bounds demonstrate that BIQ achieves a tighter error bound than SQ, resulting in more accurate quantized values. To validate these theoretical findings, we conduct sampling experiments with various quantizers. We show sampling results for uniform, bell-shaped and long-tail data distributions. As shown in Section **Sampling** (please refer to Other Experiments in supplementary materials), the empirical results align closely with our theoretical analysis, confirming the superior performance of BIQ and WBIQ in practical settings.

Convergence Analysis

BIQ is a biased quantizer, thus failing to satisfy the unbiased property required by most convergence analysis frameworks (Elgabli et al. 2020; Reisizadeh et al. 2020). To address this challenge, we develop a novel convergence analysis for biased quantizers under both strongly convex and non-convex settings. Proofs of both theorems are provided in supplementary materials. Our analysis builds on the following assumptions.

Assumption 1. Let $Q(\cdot)$ be a quantizer such that $\mathbb{E}[Q(\mathbf{x})|\mathbf{x}] = \mathbf{x} + \eta, \eta \in \mathbb{R}^d, \|\eta\| \leq G, G > 0$. The quantization variance is bounded as follows: $\mathbb{E}[\|Q(\mathbf{x}) - \mathbb{E}[Q(\mathbf{x})]\|^2|\mathbf{x}] \leq q\|\mathbf{x}\|^2, q > 0$.

Assumption 2. f_i is L-smooth if there exists a constant $L > 0$ such that $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$.

Assumption 3. The stochastic gradient $\hat{\nabla} f_i(x)$ is unbiased, i.e. $\mathbb{E}_\zeta[\hat{\nabla} f_i(\mathbf{x})] = \nabla f_i(\mathbf{x})$ with bounded variance, i.e. $\mathbb{E}_\zeta[\|\hat{\nabla} f_i(\mathbf{x}) - \nabla f_i(\mathbf{x})\|^2] \leq \sigma^2$.

Assumption 4. f_i is μ-strongly convex if there exists a constant $\mu > 0$ such that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\langle \nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu\|\mathbf{x} - \mathbf{y}\|^2$.

Theorem 1 (Convergence of Strongly Convex Settings)**.** *Suppose Assumptions 1 to 4 hold. Define* $A_0 := \frac{2qL}{\mu} + \frac{4L^2(N-s)(q+4)}{s(N-1)\mu}$. *Set a constant* k_0 *that satisfies* $k_0 \geq \max\{\frac{48\tau - 1}{\tau}, \frac{48N - \mu^2\tau}{\mu^2\tau^2}, \frac{48L^2A_0\tau - \mu^3\tau + 6L^2\mu}{\mu^2\tau^2 - L^2\mu\tau}, \frac{4L^2 - \mu^2}{\mu^2\tau}, \frac{4L\sqrt{\tau(\tau-1)} - \mu}{\mu\tau}\}$. *Then for any* $k \geq k_0$, *if the learning rate satisfies* $\alpha_k = \min\{\frac{\mu}{L^2}, \frac{1}{L\sqrt{\tau(\tau-1)}}, \frac{1}{\mu\tau}\}$, *Algorithm 2 holds*

$$\mathbb{E}\|\theta^k - \theta^*\|^2 \leq (\frac{k_0\tau + 1}{k\tau + 1})^2 \mathbb{E}\|\theta^{k_0} - \theta^*\|^2 + A_2 \frac{1}{\tau} \cdot \frac{\tau}{k\tau + 1} + A_3 \cdot \frac{\tau}{k\tau + 1} + A_4 \frac{g^{k^2}}{\tau^2} \cdot \frac{\tau}{k\tau + 1}. \tag{4}$$

where $\{g^k\}$*is a bounded positive sequence. The constants in Theorem 1 are defined as*

$$A_2 := \frac{32\sigma^2}{\mu^2}\left[\frac{e}{N}(L^2 + 1 + 2qN) + \frac{4e(N-s)(q+4)}{s(N-1)}\right],$$
$$A_3 := \frac{32\sigma^2}{\mu^2}\left[\frac{1}{N} + 2q + \frac{4(N-s)(q+4)}{s(N-1)}\right],$$
$$A_4 := \frac{3sN + 12N - 15s}{s(N-1)}.$$

Theorem 1 establishes convergence guarantees under strongly convex assumption. The bias η introduced by the biased quantizer affects the convergence rate through its influence on g^k. In FedBIQ, by choosing the quantization range as $R_n^k \leq (\frac{12}{d} \cdot 2^{2b}\|\theta_n^{k,\tau} - \theta_n^k\|^2)^{\frac{1}{2}}$, we ensure that $G_n^k \leq \frac{R_n^k}{2^b}$. Let us find $g^k = \max_{n\in[N]}\{(k\tau + 1)G_n^k\}$. This implies that the sequence $g^k$is bounded, thereby ensuring that FedBIQ achieves a convergence rate of $\mathcal{O}\left(\frac{1}{T}\right)$. For FedWBIQ, we choose the quantization range as $R_n^k \leq (\frac{48}{d} \cdot 2^{2b}\|\theta_n^{k,\tau} - \theta_n^k\|^2)^{\frac{1}{2}}$ to reach the same convergence rate.

Theorem 2 (Convergence of Non-convex Settings)**.** *Under Assumptions 1 to 3 and the update method of Algorithm 2, we define the following two constants:* $F_1 = 2L^2\tau(\tau - 1), F_2 = L + \frac{N+q+3}{N}\tau L + \frac{4L(N-s)}{s(N-1)}(q+4)\tau$. *When the constraint* $T \geq \frac{F_1^2}{(\sqrt{F_2^2 + 2F_1} - F_2)^2}$ *is satisfied and the learning rate is* $\alpha_k = \frac{1}{\sqrt{T}}$, *the following first-order stability condition holds*

$$\frac{1}{T}\sum_{k=0}^{K-1}\sum_{t=0}^{\tau-1}\mathbb{E}\|\nabla f(\bar{\theta}^{k,t})\|^2 \leq \frac{2[f(\theta^0) - f^*]}{\sqrt{T}} + H_1\frac{\tau - 1}{T} + H_2\frac{\tau}{\sqrt{T}} + H_3\frac{1}{\sqrt{T}}. \tag{5}$$

where $g^k \leq \max_{n\in[N]}\{G_n^k\}, \|\eta_n^k\| \leq G_n^k$,

$$H_1 := \frac{L\sigma^2(N+1)}{N},$$
$$H_2 := L\sigma^2\left[\frac{(N+q+3)}{N} + \frac{4(N-s)}{s(N-1)}(q+4) + \frac{1}{N\tau}\right],$$
$$H_3 := \sum_{k=0}^{K-1}[\frac{12L(N-s)}{s(N-1)}g^k + \frac{(3N+1)L}{N}g^k + \|\nabla f(\bar{\theta}^{k,\tau})\|]g^k.$$

Theorem 2 provides convergence guarantees under non-convex assumption. In FedBIQ, to satisfy the condition $T \geq \frac{F_1^2}{(\sqrt{F_2^2+2F_1}-F_2)^2}$, we select the quantization range as $R_n^k \leq \{12 \cdot 2^{2b} \cdot \frac{1}{d}[\tau L + \frac{4L(N-s)\tau}{S(N-1)}]^{-1}[\sqrt{T} - \frac{1}{2\sqrt{T}}F_1 - L - \frac{N+3}{N}\tau L - \frac{16L(N-s)\tau}{s(N-1)}]\|\theta_n^{k,r} - \theta_n^k\|^2\}^{\frac{1}{2}}$, which ensures that $G_n^k \leq \frac{R_n^k}{2^b}$. Let us find $g^k = \max_{n\in[N]}\{G_n^k\}$. This implies that g^k is bounded. We set $\frac{1}{2}R_n^k$ as the quantization range for FedWBIQ. FedBIQ and FedWBIQ converges at a rate of $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$. Notably, when employing biased quantizers, the convergence rate can be preserved as long as the bias is appropriately controlled. Moreover, if the quantizer is unbiased, i.e. $\eta = 0$, then $g^k = 0, \forall k$. The results in Theorems 1 and 2 recover the conclusions in (Reisizadeh et al. 2020), showing that our analysis generalizes existing theory.

Experiments

In this section, we conducted extensive experiments on BIQ and WBIQ. First, we compare six algorithms on four datasets, demonstrating the superior performance of BIQ in terms of cost, communication efficiency and utility. Finally, we perform quantizer sampling experiments to evaluate the accuracy of different quantization methods under low-resolution settings in supplementary materials.

Data and Models. Our experiments are conducted in a computing environment equipped with an AMD Ryzen 7 9700X CPU and an RTX 4070 Ti Super GPU. We evaluate our methods on four benchmark datasets: CDC Diabetes Health Indicators (abbreviated as CDC for convenience) (abbreviated as CDC for convenience), MNIST (Lecun et al. 1998), CIFAR-10 (Krizhevsky, Nair, and Hinton 2009) and Tiny ImageNet. CDC contains 253,680 samples consisting of 21 features and 2 labels, 70% of which are divided for training and the rest for evaluation. MNIST consists of 60,000 grayscale training images of size 28*28 pixels and 10,000 test samples, while CIFAR-10 contains 50,000 images of size 32*32 pixels across three channels, along with an equal-sized test set. MNIST and CIFAR-10 are labeled over 10 classes. To assess the practical effectiveness of BIQ on real-world datasets, we conducted comprehensive experiments on Tiny ImageNet, which comprises 100K training and 10K test images spanning 200 classes of 64×64 RGB images.

To study the impact of data heterogeneity, we adopt two data partitioning strategies: (1) I.I.D. setting: The training data is uniformly distributed across clients, ensuring balanced class distributions; (2) Non-I.I.D. setting: Data is partitioned in a heterogeneous manner using a Dirichlet distribution with concentration parameter $\alpha = 0.6$. The global test set is centrally maintained by the server to ensure a consistent and standardized evaluation of model performance across all experimental configurations.

For strongly convex settings, we build logistic regression models to solve the binary classification problem on CDC. For non-convex settings, we employ a two-layer CNN model on MNIST and CIFAR-10. Each layer of the CNN model consists of convolutional layers, ReLU activation, and max-pooling, followed by two fully connected layers. For Tiny ImageNet, we employed Resnet50.

	CDC	MNIST	CIFAR-10	Tiny ImageNet
CR	30	30	2000	50
LU	15	15	15	30
LR	0.01	0.03	0.05	0.007
MO	0.5	0.5	0.5	0.0002
BS	32	32	32	128
OP	SGD	SGD	SGD	Adam

Table 1: Experiment Settings. CR: communication rounds; LU: the number of rounds in a local update; LR: learning rate; MO: momentum; BS: batch size; OP: optimizer.

Baselines. Our work conducts a systematic comparison among BIQ, WBIQ and existing quantization-based FL methods. The baseline algorithms include: FedAvg (McMahan et al. 2017), FedSQ which incorporates stochastic quantization (Gupta et al. 2015), FedRQ which applies rounding quantization (Gupta et al. 2015), FedPAQ (Reisizadeh et al. 2020) which is based on QSGD quantization (Alistarh et al. 2017), FedNQFL (Chen et al. 2023) which incorporates nonuniform quantization, FedSQFL (Marnissi, El Hammouti, and Bergou 2024) which integrates quantization and sparsification.

Criterion. In this experiment, we establish a multidimensional evaluation framework for BIQ and WBIQ. First, we develop a cost model to quantify algorithmic efficiency. We use the shifted-exponential model (Reisizadeh et al. 2020; Lee et al. 2018) to fit the client-side encode and gradient computation time as the local computation cost. Communication cost is defined as the ratio of the total number of bits received by the server per round to the available bandwidth (BW), where BW is set to $1.5MB/s$ for CDC and MNIST, and $3MB/s$ for CIFAR-10 and Tiny ImageNet. The total cost combines both local computation and communication costs to represent the overall iteration overhead. For utility assessment, we focus on global model performance, evaluating convergence and generalization through two metrics: cross-entropy loss and classification accuracy on the test set. This dual-metric approach provides a comprehensive measure of model effectiveness. Our experiments assess algorithm optimization performance from both resource efficiency and model utility perspectives.

Numerical Results and Discussions

End-to-end Comparison. We conducted experiments on CDC, MNIST, CIFAR-10 and Tiny ImageNet. In each round, 15 clients were randomly sampled from a total of 80 for training. The specific settings are presented in Table 1.

FedAvg was implemented using 32-bit floating-point precision, while other methods were fixed to a quantization resolution of $b = 3$ bits. Figure 2 shows the variation of test loss ($Loss$) and accuracy (Acc) with the total cost ($Time$) on CDC with 1-sigma error bars, averaged over 5 runs. Table 2 report the average training cost per communication round

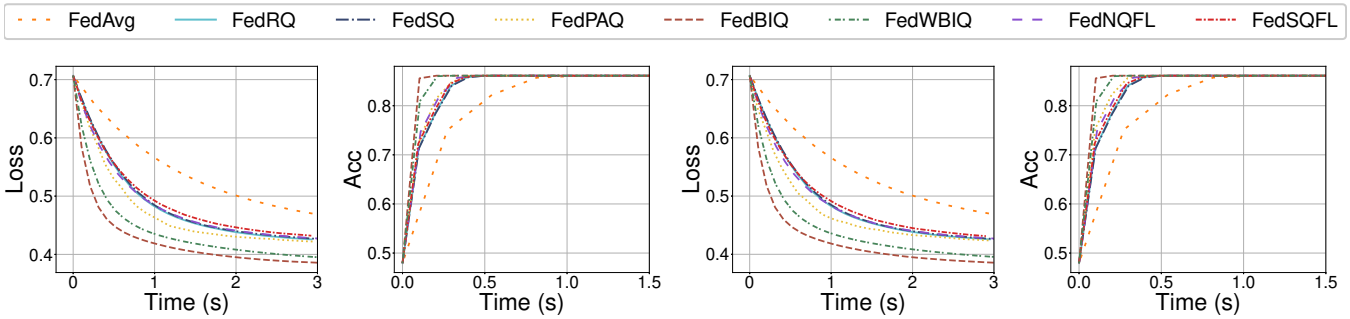

Figure 2: Loss and classification accuracy evolution on CDC under I.I.D. (left) and Non-I.I.D. (right) data distributions.

Dataset	Algorithm	I.I.D. Distribution			Non-I.I.D. Distribution		
		$cost_{avg}$	$Loss$	Acc (%)	$cost_{avg}$	$Loss$	Acc (%)
MNIST	FedAvg	100.06±3e-3	0.15±6e-4	95.58±1e-3	100.06±2e-3	0.17±2e-3	94.86±9e-4
	FedRQ	9.46±2e-3	0.38±1e-3	90.32±1e-3	9.46±1e-3	0.58±9e-3	86.82±8e-3
	FedSQ	9.47±4e-3	0.38±1e-3	90.30±3e-4	9.47±1e-3	0.57±6e-3	87.10±5e-3
	FedPAQ	9.47±1e-3	0.31±1e-2	91.23±5e-3	9.47±2e-3	0.43±1e-2	87.57±8e-3
	FedNQFL	12.05±1e-3	0.17±3e-3	94.96±1e-3	11.80±3e-3	0.20±4e-3	94.14±2e-3
	FedSQFL	**5.73±2e-3**	0.30±1e-2	91.00±5e-3	**5.73±2e-3**	0.35±2e-2	89.85±6e-3
	FedBIQ	9.47±3e-3	0.16±1e-3	95.22±7e-4	9.47±2e-3	0.19±2e-3	94.39±2e-3
	FedWBIQ	9.48±3e-3	**0.16±3e-3**	**95.37±7e-4**	9.48±2e-3	**0.18±2e-3**	**94.58±4e-4**
CIFAR-10	FedAvg	615.00±9e-4	0.63±7e-3	78.95±4e-3	615.00±1e-3	0.70±6e-3	76.51±5e-3
	FedRQ	57.76±2e-3	0.88±4e-3	69.86±2e-3	57.76±9e-4	0.95±3e-3	67.03±1e-3
	FedSQ	57.77±8e-4	0.85±3e-3	70.67±2e-3	57.76±2e-3	0.93±4e-3	67.96±2e-3
	FedPAQ	57.76±1e-3	0.85±2e-3	70.83±2e-3	57.77±3e-3	0.93±3e-3	67.96±9e-4
	FedNQFL	60.29±3e-4	0.68±4e-3	77.05±3e-3	60.09±4e-4	0.75±4e-3	74.48±3e-3
	FedSQFL	**23.96±1e-3**	0.97±3e-3	66.56±1e-3	**23.96±8e-4**	1.03±3e-3	63.68±2e-3
	FedBIQ	57.76±3e-4	0.67±1e-2	77.49±2e-3	57.77±1e-3	0.75±1e-2	74.58±3e-3
	FedWBIQ	57.77±3e-3	**0.65±5e-3**	**78.06±2e-3**	57.77±9e-4	**0.72±1e-2**	**75.56±4e-3**

Table 2: End-to-end comparison results on MNIST and CIFAR-10.

($cost_{avg}$), $Loss$ and Acc over 5 independent runs. For Tiny ImageNet, we report the top-1 accuracy (Acc_1) and top-5 accuracy (Acc_5) in Table 3. All of the results are presented in the form of mean ± standard deviation. Our supplementary materials presents the results of the Wilcoxon signed-rank test at a significance level of 0.05.

Figure 2 presents the experimental results under strongly convex settings. It illustrates that FedBIQ and FedWBIQ significantly outperform other FL algorithms in terms of convergence speed for loss reduction and accuracy improvement, across both I.I.D. and non-I.I.D. data distributions. Specifically, these methods demonstrate accelerated convergence, achieving lower loss and higher accuracy more rapidly than other methods. For non-convex settings, both FedBIQ and FedWBIQ achieve significantly lower loss (0.16) under the I.I.D. distribution on MNIST, with FedWBIQ's accuracy being only 0.21% lower than that of FedAvg. Under non-I.I.D. distribution, FedWBIQ maintain the lowest loss (0.18) and high accuracy (94.58%), indicating the robustness of the quantizer. On CIFAR-10 and Tiny ImageNet, FedBIQ and FedWBIQ also exhibit markedly higher accuracy compared to other quantization algorithms, further validating their effectiveness in more complex scenarios.

FedSQFL combines quantization with sparsification, substantially reducing training overhead; however, the dual sources of noise introduced by both techniques degrade model performance. FedNQFL introduces additional parameter overhead and computational cost, resulting in approximately 27% higher training overhead compared to FedBIQ and FedWBIQ while achieving comparable model accuracy on MNIST. FedNQFL relies on the assumption that the distribution of local gradient vectors tend to a Gaussian distribution, which is a potential factor contributing to its lower accuracy on Tiny ImageNet.

Although FedBIQ and FedWBIQ incur slightly higher training costs compared to other quantization methods due to the additional communication overhead of transmitting the quantization range, they still achieve effective cost control. The communication costs of FedBIQ and FedWBIQ remain significantly lower than those of unquantized base-

Dataset	Algorithm	I.I.D. Distribution			Non-I.I.D. Distribution		
		$cost_{avg}$	Acc_1 (%)	Acc_5 (%)	$cost_{avg}$	Acc_1 (%)	Acc_5 (%)
Tiny ImageNet	FedAvg	27378.46±1e-1	21.54±5e-3	44.30±7e-3	27378.34±2e-1	21.38±7e-3	43.84±9e-3
	FedRQ	5138.81±2e-1	17.31±1e-3	39.14±4e-3	5138.89±2e-1	17.54±2e-3	39.54±3e-3
	FedSQ	5140.28±1e-1	16.96±4e-3	38.37±4e-3	5140.12±4e-1	17.09±2e-3	38.54±3e-3
	FedPAQ	5140.51±3e-1	18.19±2e-3	40.69±4e-3	5140.24±3e-2	18.45±3e-3	41.06±3e-3
	FedNQFL	5141.03±1e-1	5.92±4e-3	17.38±9e-3	5141.17±2e-1	6.07±2e-3	17.65±5e-3
	FedSQFL	**1717.27±6e-2**	14.31±1e-3	33.51±2e-3	**1717.30±1e-1**	14.45±2e-3	33.67±2e-3
	FedBIQ	5139.31±1e-2	20.29±2e-3	43.11±5e-3	5139.51±3e-1	**20.38±3e-3**	**43.30±1e-3**
	FedWBIQ	5140.19±1e-2	**20.39±2e-3**	**43.11±4e-3**	5140.12±4e-1	20.32±8e-4	43.22±2e-3

Table 3: End-to-end comparison results on Tiny ImageNet.

lines, and the increase over other quantization methods is no more than 0.21%. Remarkably, FedBIQ and FedWBIQ achieve superior accuracy under low-bit settings, rendering the proposed quantizers highly competitive for communication-constrained FL systems.

FedBIQ vs. FedWBIQ. Theoretically, FedWBIQ admits an error upper bound that is twice that of FedBIQ. However, in practice, the two methods exhibit comparable empirical performance. This discrepancy arises from differences in the input data distribution. Our experiments reveal that input values tend to concentrate near the interval boundaries when quantization range is small. In such cases, FedWBIQ's outputs are biased toward the endpoints, yielding more accurate quantization compared to FedBIQ. Consequently, FedWBIQ demonstrates superior performance when using a larger learning rate and a smaller quantization range.

Related Work

Communication Compression in FL. Addressing the communication bottleneck in FL has become a central focus of recent research. Prior work has explored various techniques to improve communication efficiency in distributed training, including sparsification (Aji and Heafield 2017; Lin et al. 2017), periodic aggregation (McMahan et al. 2017; Sun et al. 2022) and quantization (Alistarh et al. 2017; Yue et al. 2021; Bernstein et al. 2018). While sparsification reduces communication bits by transmitting only significant updates, most deterministic sparsification schemes suffer from inadequate performance guarantees (Sun et al. 2022). Periodic aggregation reduces the frequency of communication but does not decrease the per-round transmission cost. Quantization, which represents model updates using low-precision data, provides a scalable solution by significantly compressing communication bits. This approach has proven particularly effective in resource-constrained environments, such as wireless sensor networks (Msechu and Giannakis 2012).

Quantization. Quantization can be broadly categorized into uniform and non-uniform techniques. Uniform quantization, which includes RQ (Sun et al. 2022; Nagel et al. 2020; Lee, Kim, and Ham 2021; Nagel et al. 2022) and SQ (Elgabli et al. 2020; Reisizadeh et al. 2020; Alistarh et al. 2017; Tran et al. 2019), utilizes evenly spaced intervals, providing simplicity but often struggling to effectively represent non-uniform distributions. In contrast, non-uniform quantization (Zhang et al. 2018; Jung et al. 2019; Gongyo et al. 2024; Li, Dong, and Wang 2019) dynamically allocates intervals based on distributions, thereby enhancing accuracy. However, its implementation in FL has been limited due to the challenges of aligning quantization schemes across clients without incurring extra communication costs. Recent hybrid approaches (Feng and Venkitasubramaniam 2024) and privacy-focused variants (Feng and Venkitasubramaniam 2025; Youn et al. 2023) have investigated trade-offs between accuracy and privacy.

Theoretical Guarantees. Prior work (Alistarh et al. 2017) analyzed the convergence of unbiased quantizers in distributed learning, focusing on efficient computation on GPUs rather than FL. Existing convergence analyses for quantized FL, such as those by (Elgabli et al. 2020) and (Reisizadeh et al. 2020), assume unbiased quantization errors. The assumption limits their applicability to biased quantizers. Prior work has argued that directly applying biased compression in FL leads to non-convergence (Li and Li 2023). Our work bridges this gap by dynamically controlling the quantization range and establishing the first convergence guarantees for biased quantizers under both strongly convex and non-convex settings. We generalizing prior theoretical framework and demonstrate the feasibility of high-precision, communication-efficient quantization in FL systems.

Conclusion

In our work, we address the communication bottleneck in FL by proposing two efficient non-uniform quantizers: BIQ and WBIQ. BIQ leverages a bisection-based encoding mechanism to achieve higher quantization accuracy without increasing communication costs. WBIQ further enhances robustness by improving the quantization of boundary values through maximum likelihood estimation. Additionally, we provide the first convergence analysis for biased quantizers within the FL framework, thereby extending theoretical guarantees beyond the limitations of unbiased quantization. Our experiments validate that BIQ and WBIQ outperform existing methods under both I.I.D. and non-I.I.D. distributions. The end-to-end training cost of BIQ and WBIQ is comparable to that of other quantization algorithms, while the model accuracy remains close to that of their full-precision counterparts. Our work offers a practical and theoretically grounded solution for resource-constrained FL, ensuring efficient communication and robust performance.

References

Aji, A. F.; and Heafield, K. 2017. Sparse communication for distributed gradient descent. *arXiv preprint arXiv:1704.05021*.

Alistarh, D.; Grubic, D.; Li, J.; Tomioka, R.; and Vojnovic, M. 2017. QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in neural information processing systems*, 30.

Bai, Y.; Wang, Y.-X.; and Liberty, E. 2018. Proxquant: Quantized neural networks via proximal operators. *arXiv preprint arXiv:1810.00861*.

Bernstein, J.; Wang, Y.-X.; Azizzadenesheli, K.; and Anandkumar, A. 2018. signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, 560–569. PMLR.

Boscarino, N.; Cartwright, R. A.; Fox, K.; and Tsosie, K. S. 2022. Federated learning and Indigenous genomic data sovereignty. *Nature machine intelligence*, 4(11): 909–911.

Chen, G.; Xie, K.; Tu, Y.; Song, T.; Xu, Y.; Hu, J.; and Xin, L. 2023. Nqfl: Nonuniform quantization for communication efficient federated learning. *IEEE Communications Letters*, 28(2): 332–336.

Cui, L.; Qu, Y.; Xie, G.; Zeng, D.; Li, R.; Shen, S.; and Yu, S. 2021. Security and privacy-enhanced federated learning for anomaly detection in IoT infrastructures. *IEEE Transactions on Industrial Informatics*, 18(5): 3492–3500.

Elgabli, A.; Park, J.; Bedi, A. S.; Bennis, M.; and Aggarwal, V. 2020. Q-GADMM: Quantized Group ADMM for Communication Efficient Decentralized Machine Learning. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8876–8880.

Feng, C.; and Venkitasubramaniam, P. 2024. RQP-SGD: Differential Private Machine Learning through Noisy SGD and Randomized Quantization. *arXiv preprint arXiv:2402.06606*.

Feng, C.; and Venkitasubramaniam, P. 2025. Randomized Quantization for Privacy in Resource Constrained Machine Learning at-the-edge and Federated Learning. *IEEE Transactions on Machine Learning in Communications and Networking*, 1–1.

Gongyo, S.; Liang, J.; Ambai, M.; Kawakami, R.; and Sato, I. 2024. Learning Non-uniform Step Sizes for Neural Network Quantization. In *Proceedings of the Asian Conference on Computer Vision*, 4385–4402.

Gupta, S.; Agrawal, A.; Gopalakrishnan, K.; and Narayanan, P. 2015. Deep learning with limited numerical precision. In *International conference on machine learning*, 1737–1746. PMLR.

Han, S.; Mao, H.; and Dally, W. J. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.

Jung, S.; Son, C.; Lee, S.; Son, J.; Han, J.-J.; Kwak, Y.; Hwang, S. J.; and Choi, C. 2019. Learning to quantize deep networks by optimizing quantization intervals with task loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4350–4359.

Krizhevsky, A.; Nair, V.; and Hinton, G. 2009. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto. CIFAR-10 dataset.

Lecun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Lee, J.; Kim, D.; and Ham, B. 2021. Network quantization with element-wise gradient scaling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6448–6457.

Lee, K.; Lam, M.; Pedarsani, R.; Papailiopoulos, D.; and Ramchandran, K. 2018. Speeding Up Distributed Machine Learning Using Codes. *IEEE Transactions on Information Theory*, 64(3): 1514–1529.

Li, X.; and Li, P. 2023. Analysis of error feedback in federated non-convex optimization with biased compression: Fast convergence and partial participation. In *International Conference on Machine Learning*, 19638–19688. PMLR.

Li, Y.; Dong, X.; and Wang, W. 2019. Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks. *arXiv preprint arXiv:1909.13144*.

Lin, Y.; Han, S.; Mao, H.; Wang, Y.; and Dally, W. J. 2017. Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*.

Luqman, A.; Qazi, K.; and Khan, I. 2024. Post-Training Non-Uniform Quantization for Convolutional Neural Networks. *arXiv preprint arXiv:2412.07391*.

Marnissi, O.; El Hammouti, H.; and Bergou, E. H. 2024. Adaptive sparsification and quantization for enhanced energy efficiency in federated learning. *IEEE Open Journal of the Communications Society*.

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

Miyashita, D.; Lee, E. H.; and Murmann, B. 2016. Convolutional neural networks using logarithmic data representation. *arXiv preprint arXiv:1603.01025*.

Msechu, E. J.; and Giannakis, G. B. 2012. Sensor-Centric Data Reduction for Estimation With WSNs via Censoring and Quantization. *IEEE Transactions on Signal Processing*, 60(1): 400–414.

Nagel, M.; Amjad, R. A.; Van Baalen, M.; Louizos, C.; and Blankevoort, T. 2020. Up or down? adaptive rounding for post-training quantization. In *International conference on machine learning*, 7197–7206. PMLR.

Nagel, M.; Fournarakis, M.; Bondarenko, Y.; and Blankevoort, T. 2022. Overcoming oscillations in quantization-aware training. In *International Conference on Machine Learning*, 16318–16330. PMLR.

Pati, S.; Baid, U.; Edwards, B.; Sheller, M.; Wang, S.-H.; Reina, G. A.; Foley, P.; Gruzdev, A.; Karkada, D.; Davatzikos, C.; et al. 2022. Federated learning enables big data

for rare cancer boundary detection. *Nature communications*, 13(1): 7346.

Reisizadeh, A.; Mokhtari, A.; Hassani, H.; Jadbabaie, A.; and Pedarsani, R. 2020. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International conference on artificial intelligence and statistics*, 2021–2031. PMLR.

Sun, J.; Chen, T.; Giannakis, G. B.; Yang, Q.; and Yang, Z. 2022. Lazily Aggregated Quantized Gradient Innovation for Communication-Efficient Federated Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4): 2031–2044.

Tran, N. H.; Bao, W.; Zomaya, A.; Nguyen, M. N.; and Hong, C. S. 2019. Federated learning over wireless networks: Optimization model design and analysis. In *IEEE INFOCOM 2019-IEEE conference on computer communications*, 1387–1395. IEEE.

Wang, L.; Dong, X.; Wang, Y.; Liu, L.; An, W.; and Guo, Y. 2022. Learnable lookup table for neural network quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12423–12433.

Youn, Y.; Hu, Z.; Ziani, J.; and Abernethy, J. 2023. Randomized quantization is all you need for differential privacy in federated learning. *arXiv preprint arXiv:2306.11913*.

Yue, L.; Cai, Y.; Zhu, M.; Wang, P.; Zhang, L.; Sun, M.; Liang, S.; Lei, M.; Zhang, J.; Hua, B.; Tian, L.; Zou, Y.; and Li, A. 2021. Improving Performance of Direct-Detection Terahertz Communication System based on k-Means Adaptive Vector Quantization. In *2021 19th International Conference on Optical Communications and Networks (ICOCN)*, 1–3.

Zhang, D.; Yang, J.; Ye, D.; and Hua, G. 2018. Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, 365–382.