Communication-Efficient Federated Learning with Adaptive Quantization

YUZHU MAO and ZIHAO ZHAO, Tsinghua Shenzhen International Graduate School, Tsinghua University, China and Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University, China
GUANGFENG YAN, City University of Hong Kong, Hong Kong, China and City University of Hong Kong Shenzhen Research Institute, China
YANG LIU, Institute for AI Industry Research, Tsinghua University, China
TIAN LAN, George Washington University, USA
LINQI SONG, City University of Hong Kong, Hong Kong, China and City University of Hong Kong Shenzhen Research Institute, China
WENBO DING, Tsinghua Shenzhen International Graduate School, Tsinghua University, China, Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University, China, and RISC-V International Open Source Laboratory, China

Federated learning (FL) has attracted tremendous attentions in recent years due to its privacy-preserving measures and great potential in some distributed but privacy-sensitive applications, such as finance and health. However, high communication overloads for transmitting high-dimensional networks and extra security masks remain a bottleneck of FL. This article proposes a communication-efficient FL framework with an Adaptive Quantized Gradient (AQG), which adaptively adjusts the quantization level based on a local gradient's update to fully utilize the heterogeneity of local data distribution for reducing unnecessary transmissions. In addition, client dropout issues are taken into account and an Augmented AQG is developed, which could limit the dropout noise with an appropriate amplification mechanism for transmitted gradients. Theoretical analysis and experiment results show that the proposed AQG leads to 18% to 50% of additional transmission reduction as compared with existing popular methods, including Quantized Gradient Descent

This work is supported in part by the National Natural Science Foundation of China under grant no. 62104125, by the Guangdong Basic and Applied Basic Research Foundation under grant no. 2020A1515110887, by the Institute for Guo Qiang of Tsinghua University under grant no. 2020GQG1004, by Tsinghua Shenzhen International Graduate School under grant nos. HW202 and QD2021013C, and by the Hong Kong RGC under grant no. ECS 21212419.
Authors' addresses: Y. Mao and Z. Zhao, Tsinghua Shenzhen International Graduate School, Tsinghua University, Lishui Rd, Shenzhen, Guangdong Province, China, 518055, Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University, Lishui Rd, Shenzhen, Guangdong Province, China, 518055; emails: {myz20, zhao-zh21}@mails.tsinghua.edu.cn; G. Yan and L. Song, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong, China, City University of Hong Kong Shenzhen Research Institute, Yuexing First Rd, Shenzhen, Guangdong Province, China, 518057; emails: gfyan2-c@my.cityu.edu.hk, linqi.song@cityu.edu.hk; Y. Liu, Institute for AI Industry Research, Tsinghua University, Shuangqing Rd, Beijing, China, 100190; email: liuy03@air.tsinghua.edu.cn; T. Lan, Department of Electrical and Computer Engineering, George Washington University, 1918 F Street, NW, Washington, DC, USA, 20052; email: tlan@gwu.edu; W. Ding (corresponding author), Tsinghua Shenzhen International Graduate School, Tsinghua University, Lishui Rd, Shenzhen, Guangdong Province, China, 518055, Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University, Lishui Rd, Shenzhen, Guangdong Province, China, 518055, RISC-V International Open Source Laboratory, Lishui Rd, Shenzhen, Guangdong Province, China, 518055; email: ding.wenbo@sz.tsinghua.edu.cn.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Association for Computing Machinery.
2157-6904/2022/08-ART67 $15.00
https://doi.org/10.1145/3510587

(QGD) and Lazily Aggregated Quantized (LAQ) gradient-based methods without deteriorating convergence properties. Experiments with heterogenous data distributions corroborate a more significant transmission reduction compared with independent identical data distributions. The proposed AQG is robust to a client dropping rate up to 90% empirically, and the Augmented AQG manages to further improve the FL system's communication efficiency with the presence of moderate-scale client dropouts commonly seen in practical FL scenarios.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Machine learning**; **Distributed computing methodologies**; • **Security and privacy**; • **Networks** → *Network reliability*;;

Additional Key Words and Phrases: Federated learning, information compression, communication efficiency

ACM Reference format:
Yuzhu Mao, Zihao Zhao, Guangfeng Yan, Yang Liu, Tian Lan, Linqi Song, and Wenbo Ding. 2022. Communication-Efficient Federated Learning with Adaptive Quantization. *ACM Trans. Intell. Syst. Technol.* 13, 4, Article 67 (August 2022), 26 pages.
https://doi.org/10.1145/3510587

1 INTRODUCTION

The deployment of the Internet of things (IoT), ubiquitous sensing, edge computing, and many other distributed systems have enabled the rapid development of distributed learning techniques in recent years [10, 12, 20]. Distributed learning could fully utilize low-cost computing resources throughout the network and achieve comparable performance with centralized learning. Nevertheless, the leakage of data, gradients, and even models during the updating and transmitting process in distributed learning has raised the concerns of user privacy and security, which greatly limit its applications in some specific fields, such as finance and health. To this end, federated learning (FL), which prevents privacy leakage by avoiding data exposition, has been proposed by Google and other researchers, attracting tremendous attention from both academia and industry [22].

Many approaches — such as differential privacy [1], secret-sharing techniques [5], and homomorphic encryption [21] — have been developed to mask transmitted gradients and can mostly well address the security issues in FL. However, high-dimensional neural networks and extra security masks [8, 16, 31] may lead to high communication overhead, which becomes a main bottleneck of FL systems. In this context, communication-efficient learning algorithms have been proposed mainly to reduce transmission bits based on gradient quantization, which maps a real-valued vector to a constant number of bits. Representative gradient quantization algorithms for distributed systems include Quantized Stochastic Gradient Descent (QSGD) [3], 1-bit SGD [25], and SignSGD [4]. However, these methods communicate at all iterations (transmit all computed gradients) with a fixed number of quantization bits, which is not efficient enough for FL, in which non-IID (Independently Identically Distributed) data distribution is common. To address this problem, Sun et al. proposed a gradient innovation-based Lazily Aggregated Quantized (LAQ) gradient method, which utilizes the differences between local loss functions and skips the transmission of slowly varying quantized gradients [15]. Although the LAQ method reduces transmission overload by skipping unnecessary communication rounds, it still fixes the number of bits for all transmitted gradients, which remains to be improved.

In order to further reduce overall transmitted bits, this article proposes a communication-efficient FL framework with an **A**daptive **Q**uantized **G**radient (AQG), in which the quantization level is adjusted according to the local gradient's updates adaptively. Specifically, gradients with a larger amount of updates are quantized and transmitted with more bits and vice versa. In addition, this article takes client dropouts into account, which is another main challenge faced by FL system due to limited device reliability [5]. In order to improve the performance of an AQG with the presence of the noise introduced by client dropouts, the proposed FL framework with an

Table 1. Notations

$\boldsymbol{g}_m^k$	gradient computed by client m at iteration k
$\hat{\boldsymbol{g}}_m^k$	gradient used for aggregation from client m at iteration k
b_{max}	upper bound for the number of bits after quantization
b_m^k	the quantization bit number chosen by client m at iteration k
$\hat{b}_m^k$	the quantization bit number chosen by client m for $\hat{\boldsymbol{g}}_m^k$
$Q_b(\boldsymbol{g}_m^k)$	$\boldsymbol{g}_m^k$ quantized with b bits
$\boldsymbol{\theta}^k$	the aggregated global model broadcast at iteration k
$\varepsilon_b(\boldsymbol{g}_m^k)$	quantization error $(Q_b(\boldsymbol{g}_m^k) - \boldsymbol{g}_m^k)$
$\mathbb{M}$	client set
$\mathbb{M}_b^k$	subset of clients uploading gradients with b bits at iteration k
p	client dropping rate
$\lceil a \rceil$	the ceil of a
$\|\mathbf{x}\|_2$	l_2-norm of $\mathbf{x}$
$\|\mathbf{x}\|_\infty$	l_∞-norm of $\mathbf{x}$

AQG is augmented by a variance-reduced method in which transmitted gradients are appropriately amplified to keep the unbiased estimators.

Theoretical analysis and experiment results show that the proposed AQG outperforms existing methods in terms of overall transmitted bits without deteriorating convergence properties. The AQG is robust to a client dropping rate up to 90% empirically, and the Augmented AQG with gradient amplification acts as a competitive solution to achieve an even more significant transmission reduction, with a moderate client dropping scale commonly seen in practical FL scenarios.

The remainder of the article is organized as follows. Section 2 provides an overview of the FL system and discusses our motivations. The proposed Adaptive Quantized Gradient method is elaborated in Section 3. A theoretical analysis and convergence guarantee of AQG are provided in Section 4. We evaluate the performance of AQG with extensive experiments in Section 5 and present our conclusions in Section 6.

Notations. The notations used in this article are listed in Table 1.

2 SYSTEM OVERVIEW AND MOTIVATIONS

2.1 Federated Learning System

FL is designed to collaboratively train a global machine learning model with heterogeneous local data distribution across multiple privacy-sensitive clients. A typical architecture for a FL system with M distributed clients and a server is shown in Figure 1. Similar to most distributed learning systems, an FL system uses a server to receive locally computed gradients and update the global model by aggregation. However, in order to prevent privacy leakage from raw gradients, distributed clients have to mask or encrypt the local gradients before transmission. Therefore, the communication burden in FL systems tends to be heavier compared with other distributed learning systems [5]. In addition, distributed clients in FL systems, such as mobile devices in wireless networks, usually have limited computation and communication resources, which may lead to the dropout of the participants in each iteration, like the client M shown in Figure 1. Thus, the robustness to client dropout is another practical requirement for FL systems [5].

2.2 Motivations

FL is bottlenecked by high communication overhead and limited device reliability. The lack of efficient transmission and robustness to client dropouts may lead to slow, expensive, and unstable

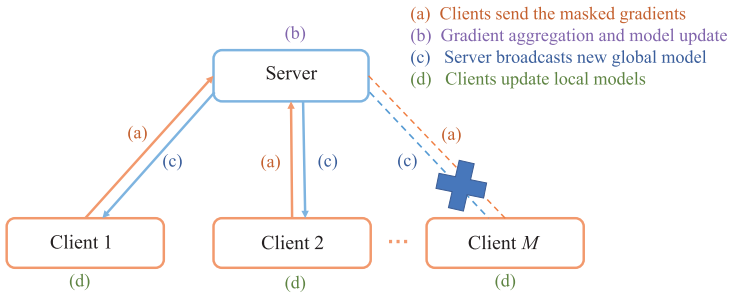


Fig. 1. Typical architecture for an FL system.

learning. In this article, the FL framework with the proposed AQG method provides opportunities for communication-efficient FL with large-scale of client dropouts.

First, AQG focuses on reducing unnecessary transmission by fully utilizing the heterogeneous property of FL. Due to the heterogeneity of local data distribution, local optimization objectives decrease at different rates. Therefore, adaptively adjusting the quantization level according to a gradient's update amount provides a more efficient way to communicate with the server by quantizing slowly varying gradients with less amount of bits.

Second, AQG aims to address the noise induced by client dropouts. When a client dropout occurs, all coordinates of a transmitted gradient are lost, which can be regarded as an extreme example of gradient sparsification [2, 19, 28, 29]. In order to limit the variance increase of a sparsified gradient, Wangni et al. proposed keeping the unbiasedness of the sparsified gradient by appropriately amplifying the remaining coordinates [30]. Inspired by this idea, the AQG tries to stay robust to client dropouts or even further improve the communication efficiency of FL with client dropouts by further adjusting the transmitted gradients and suppressing the noise.

3 AQG: ADAPTIVE QUANTIZED GRADIENT

To reduce transmission overhead, a multilevel adaptive quantization scheme is proposed in this section. As illustrated in Figure 2(a), the FL system with an AQG can be implemented as follows. At iteration k, the server broadcasts global model $\boldsymbol{\theta}^k$ to all clients. Each client computes gradient $\boldsymbol{g}_m^k$ with its local data $\mathbf{X}_m$:

$$\boldsymbol{g}_m^k = \nabla f_m(\mathbf{X}_m; \boldsymbol{\theta}^k). \tag{1}$$

After gradient computation, each client needs to make **two decisions**: (1) is it necessary to send its quantized gradient? and (2) how many bits b_m^k should be used to quantize and send its newly computed gradient? The first decision is the key idea in the LAQ method [15]. In this article, it is considered to be a special case of the second decision, where b_m^k is chosen as zero if the client decides to send nothing.

If client m chooses a non-zero b_m^k and updates its newly quantized gradient, then $Q_{b_m^k}(\boldsymbol{g}_m^k)$ is one of the quantized gradients that actually participates in gradient aggregation on the server side at iteration k. Otherwise, the server reuses the old quantized-gradient $Q_{\hat{b}_m^{k-1}}(\hat{\boldsymbol{g}}_m^{k-1})$ from the last iteration to represent client m in the aggregation. In summary, an iteration step of the proposed AQG is as follows:

Gradient Update

$$Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right) = \begin{cases} Q_{b_m^k}\left(\boldsymbol{g}_m^k\right), & m \in \mathbb{M} \setminus \mathbb{M}_0^k \\ Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right), & m \in \mathbb{M}_0^k \end{cases} \tag{2}$$

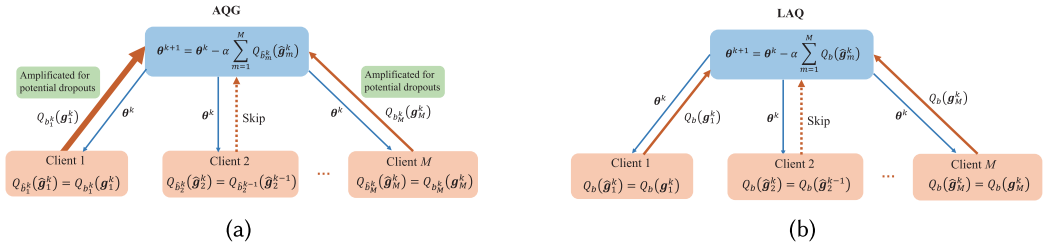


Fig. 2. The schematic illustration of the communication-efficient FL with an AQG in comparison with the LAQ method In LAQ, the quantization level is fixed at b, while the AQG adaptively adjusts the quantization level for every client at each iteration, as indicated by (a), in which the red lines indicating the transmission of quantized gradients are drawn in different thicknesses to represent different quantization levels selected by various clients. In addition, the AQG addresses potential client dropouts with appropriate gradient amplification.

Gradient Aggregation

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \alpha \sum_{m \in \mathbb{M}} Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right), \tag{3}$$

where $\mathbb{M}_0^k$ denotes the subset of clients that sets $b_m^k = 0$ and uploads nothing at iteration k. For client m, $Q_{\hat{b}_m^k}(\hat{\boldsymbol{g}}_m^k)$ represents the quantized gradient actually used for aggregation at iteration k, which may be outdated if $m \in \mathbb{M}_0^k$.

The target problems of AQG are that:

(1) For adaptive quantization of lazily aggregated gradients, a **precision selection criterion** that can cooperate with a lazy aggregation scheme and adaptively decide the quantization level for each newly computed gradient is required.
(2) For an FL scenario in which client dropouts are relatively frequent, methods to limit the noise introduced by gradient lossing are also in great need.

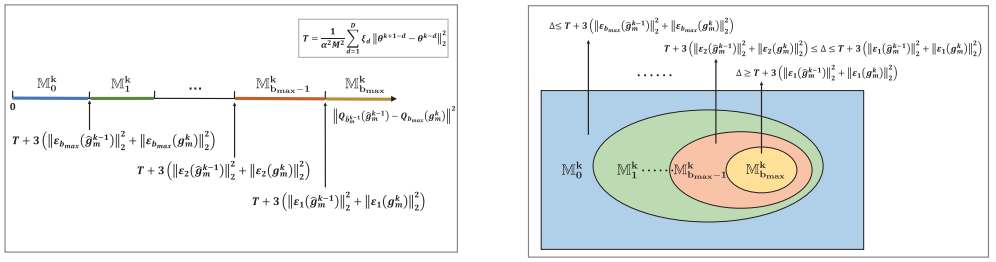
The next section presents the precision selection criterion developed in this article and the quantization scheme applied in the proposed AQG. In Section 3.3, an optional augmentation of AQG is proposed to address potential client dropouts.

3.1 Precision Selection Criterion

As mentioned before, the LAQ algorithm proposed by Sun et al. skips the uploads of quantized gradients with small innovations — the difference between $Q_b(\boldsymbol{g}_m^k)$ and the last upload $Q_b(\hat{\boldsymbol{g}}_m^{k-1})$, where b is the fixed number of bits after quantization [15]. In order to decide whether client m needs to upload its newly quantized gradient $Q_b(\boldsymbol{g}_m^k)$ at iteration k, the LAQ method develops a communication selection criterion as follows:

$$\left\|Q_b\left(\hat{\boldsymbol{g}}_m^{k-1}\right) - Q_b\left(\boldsymbol{g}_m^k\right)\right\|_2^2 \geq \frac{1}{\alpha^2 M^2} \sum_{d=1}^{D} \xi_d \|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2 + 3\left(\left\|\varepsilon_b\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_2^2 + \|\varepsilon_b\left(\boldsymbol{g}_m^k\right)\|_2^2\right), \tag{4}$$

where $\varepsilon_b(\hat{\boldsymbol{g}}_m^{k-1})$ and $\varepsilon_b(\boldsymbol{g}_m^k)$ denote quantization errors, and $\{\xi_d\}_{d=1}^D$ are predetermined constant weights used to balance the impact of global model updates from previous D steps. In LAQ, client m sends its newly quantized local gradient $Q_b(\boldsymbol{g}_m^k)$ at iteration k only when the difference between $Q_b(\boldsymbol{g}_m^k)$ and the last upload $Q_b(\hat{\boldsymbol{g}}_m^{k-1})$ is larger than a threshold, which takes the quantization error and global model's innovation into account [15]. Note that the quantization level b in LAQ is fixed.



(a) (b)

Fig. 3. The principle of the precision selection criterion, in which T denotes the parameter difference term $\frac{1}{\alpha^2 M^2}\sum_{d=1}^{D}\xi_d\|\boldsymbol{\theta}^{k+1-d}-\boldsymbol{\theta}^{k-d}\|_2^2$ and Δ denotes the gradient innovation $\|Q_{\hat{b}_m^{k-1}}(\hat{\boldsymbol{g}}_m^{k-1})-Q_{b_{max}}(\boldsymbol{g}_m^k)\|_2^2$.

This article extends the single precision level LAQ with communication selection criterion (4) to multilevel adaptive quantization for transmitted gradients. The key idea of the AQG is that under a preset upper bound b_{max} for the number of bits after quantization, gradients with smaller innovations can be quantized with a lower number of bits, since the negative impact of their precision losses on convergence is limited.

In order to decide how many bits b_m^k should be used to quantize and send client m's newly computed gradient $\boldsymbol{g}_m^k$, we develop the following precision selection criterion:

$$\begin{aligned}&\left\|Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)-Q_{b_{max}}\left(\boldsymbol{g}_m^k\right)\right\|_2^2\\&\quad\geq\frac{1}{\alpha^2M^2}\sum_{d=1}^{D}\xi_d\|\boldsymbol{\theta}^{k+1-d}-\boldsymbol{\theta}^{k-d}\|_2^2+3\left(\left\|\varepsilon_{b_{max}-b+1}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_2^2+\left\|\varepsilon_{b_{max}-b+1}\left(\boldsymbol{g}_m^k\right)\right\|_2^2\right).\end{aligned}\tag{5}$$

As illustrated in Figure 3, the precision selection criterion (5) in the AQG quantizes larger updates with more bits and vice versa. Specifically, with quantization levels $[1,\ldots,b_{max}]$, the quantization errors for any given vector $\boldsymbol{g}$ always satisfy $\varepsilon_{b_{max}}(\boldsymbol{g})\leq\varepsilon_{b_{max}-1}(\boldsymbol{g})\cdots\leq\varepsilon_1(\boldsymbol{g})$. Therefore, with b_{max} quantization levels in total, the range of each gradient innovation $\|Q_{\hat{b}_m^{k-1}}(\hat{\boldsymbol{g}}_m^{k-1})-Q_{b_{max}}(\boldsymbol{g}_m^k)\|_2^2$ can be divided into $b_{max}+1$ intervals, as shown in Figure 3(a). Then, we allocate the higher quantization level for clients with larger gradient innovations, as indicated by Figure 3(b). Thus, the proposed precision selection criterion divides the entire client set $\mathbb{M}$ into $b_{max}+1$ non-overlapping subsets as follows:

$$\mathbb{M}_0^k\cup\mathbb{M}_1^k\cup\mathbb{M}_2^k\cup\mathbb{M}_3^k\cup\cdots\cup\mathbb{M}_{b_{max}}^k=\mathbb{M}\tag{6a}$$

$$\mathbb{M}_0^k\cap\mathbb{M}_1^k\cap\mathbb{M}_2^k\cap\mathbb{M}_3^k\cap\cdots\cap\mathbb{M}_{b_{max}}^k=\varnothing,\tag{6b}$$

where $\mathbb{M}_b^k$ denotes the subset of clients that send gradients quantized by b bits at iteration k. $\mathbb{M}_0^k$ denotes the subset of clients that skip the update.

FL with an AQG is summarized in Algorithm 1. At iteration k, each client checks where its innovation locates in Figure 3(a), and then re-quantizes its gradient with the corresponding number of bits for the update. Theoretical analysis of a multilevel AQG with (5) is provided in Section 4.

For computation simplicity, a two-level variant of the AQG is also proposed in this article. At each iteration:

Two-level AQG. There are only two precision levels to be selected for each client. In other words, b in criterion (5) only has two options: $\lceil\frac{b_{max}}{2}\rceil$ and b_{max}.

ALGORITHM 1: AQG

Input: stepsize $\alpha > 0$, b_{max}, D, and $\{\xi_d\}_{d=1}^{D}$.
Initialize: θ^1.

for $k = 1, 2, \ldots, K$ **do**
  Server broadcasts θ^k to all workers.
  for each client $m \in \mathbb{M}$ **in parallel do**
    Worker m computes g_m^k and $Q_{b_{max}}(g_m^k)$.
    if (5) with $b = 1$ holds for worker m **then**
      for $b = b_{max}, b_{max} - 1, \ldots, 1$ **do**
        if (5) with b holds for worker m **then**
          Worker m computes and sends $Q_b(g_m^k)$.
          Set $b_m^k = b$.
          Set $\hat{g}_m^k = g_m^k$ and $\hat{b}_m^k = b$ on both sides.
          Break.
        end if
      end for
    else
      Worker m sends nothing.
      Set $b_m^k = 0$,
      Set $\hat{g}_m^k = \hat{g}_m^{k-1}$ and $\hat{b}_m^k = \hat{b}_m^{k-1}$ on both sides.
    end if
  end for
  Server updates θ^{k+1} by $\theta^k - \alpha \sum_{m=1}^{M} Q_{\hat{b}_m^k}(\hat{g}_m^k)$.
end for

3.2 Quantization Scheme

For better comparison, we adapt the quantization scheme used in the LAQ algorithm [15]. The scheme quantizes the difference between the new gradient g_m^k and the last quantized upload $Q_{\hat{b}_m^{k-1}}(\hat{g}_m^{k-1})$:

$$\Delta = g_m^k - Q_{\hat{b}_m^{k-1}}\left(\hat{g}_m^{k-1}\right). \tag{7}$$

With b bits used for quantization, the value range of Δ's elements can be represented by a uniformly discretized grid with $2^b - 1$ quantized values, as shown in Figure 4. By projecting every real number in this range to the closest quantized value, g_m^k can be represented by $Q_b(g_m^k)$ with b bits for each element instead of 32/64 bits by default.

3.3 Augmented AQG for Client Dropouts

This article also considers random client dropout in FL and uses z_m^k to control the participation of client m at iteration k. With a client dropping rate p:

$$z_m^k \sim Bernoulli(p).$$

If $z_m^k = 1$, client m drops out and fails to perform gradient computation at iteration k. It is obvious that with a dropping rate p, the percentage of active clients is approximately $1 - p$ at each iteration.

With this setting, the expectation of client m's upload is as follows:

$$E\left[Q_{b_m^k}\left(g_m^k\right)\right] = (1-p) \cdot Q_{b_m^k}\left(g_m^k\right) + p \cdot \mathbf{0}, \tag{8}$$

where $\mathbf{0}$ is a zero vector of the same shape as $Q_{b_m^k}(g_m^k)$.

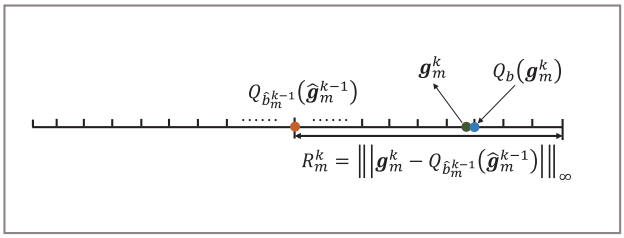


Fig. 4. Quantization scheme in the AQG.

In order to get an unbiased expectation, the upload is adjusted to $Q_{b_m^k}(\boldsymbol{g}_m^k)/(1-p)$. Then,

$$E\left[Q_{b_m^k}\left(\boldsymbol{g}_m^k\right)\right] = (1-p)\cdot\left(Q_{b_m^k}\left(\boldsymbol{g}_m^k\right)/(1-p)\right) + p\cdot\mathbf{0} = Q_{b_m^k}\left(\boldsymbol{g}_m^k\right). \tag{9}$$

The Augmented AQG is summarized in Algorithm 2. The intuitive explanation for gradient amplification is that the loss function f_m is smooth, which means the new update $Q_{b_m^k}(\boldsymbol{g}_m^k)$ tends to be approximate to recent previous updates that may have been lost due to client dropouts.

Compared with the existing LAQ method, the proposed AQG method adjusts the number of quantization bits based on local gradient innovation adaptively. The rationale of AQG is that the proposed precision selection criterion utilizes the inherent heterogeneity of local optimization objectives to reduce unnecessary transmission cost. Theoretical analysis in the next section will prove that the AQG maintains the desired convergence properties of the LAQ method. Experiments show that the AQG advances and fits FL better with the following contributions:

(1) The AQG outperforms existing popular methods in terms of overall transmission bits and achieves a more significant transmission reduction with heterogeneous data distribution compared with IID data distribution.
(2) The AQG is robust to a client's dropping rate up to 90%, and the Augmented AQG manages to further reduce transmission overload with a moderate scale of client dropouts.

4 CONVERGENCE ANALYSIS

In this section, the proposed AQG is analyzed theoretically and a convergence guarantee is provided. The theoretical analysis of an AQG is based on the following assumption:

Assumption 1. *Loss function $f(\boldsymbol{\theta}) = \sum_{m\in\mathbb{M}} f_m(\boldsymbol{\theta})$ is L-smooth.*

The Lyapunov function of AQG is defined in the same way as the LAQ:

$$\mathbb{V}(\boldsymbol{\theta}^k) = f(\boldsymbol{\theta}^k) - f(\boldsymbol{\theta}^*) + \sum_{d=1}^{D}\sum_{j=d}^{D}\frac{\xi_j}{\alpha}\|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2, \tag{10}$$

where $\boldsymbol{\theta}^*$ is the optimal solution of $\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$.

With the quantization errors in precision selection criterion (5) being ignored, the parameter differences term in Lyapunov function helps guarantee that the error induced by skipping gradients decreases with the objective residual in the training process.

4.1 Convergence Guarantee

To ensure convergence, the following inequality should always hold:

$$\mathbb{V}(\boldsymbol{\theta}^{k+1}) - \mathbb{V}(\boldsymbol{\theta}^k) \le 0. \tag{11}$$

ALGORITHM 2: Augmented AQG

Input: stepsize $\alpha > 0$, b_{max}, D, and $\{\xi_d\}_{d=1}^{D}$.
Initialize: $\boldsymbol{\theta}^1$.

for $k = 1, 2, \ldots, K$ **do**
  Server broadcasts $\boldsymbol{\theta}^k$ to all workers.
  for each client $m \in \mathbb{M}$ **in parallel do**
    if $z_m^k = 0$ **then**
      Worker m computes $\boldsymbol{g}_m^k$ and $Q_{b_{max}}(\boldsymbol{g}_m^k)$.
      if (5) with $b = 1$ holds for worker m **then**
        for $b = b_{max}, b_{max} - 1, \ldots, 1$ **do**
          if (5) with b holds for worker m **then**
            Worker m computes and sends $Q_b(\boldsymbol{g}_m^k)$.
            Set $b_m^k = b$.
            Set $\hat{\boldsymbol{g}}_m^k = \boldsymbol{g}_m^k$ and $\hat{b}_m^k = b$ on both sides.
            Break.
          end if
        end for
      end if
    else
      Worker m sends nothing.
      Set $b_m^k = 0$,
      Set $\hat{\boldsymbol{g}}_m^k = \hat{\boldsymbol{g}}_m^{k-1}$ and $\hat{b}_m^k = \hat{b}_m^{k-1}$ on both sides.
    end if
  end for
  Server updates $\boldsymbol{\theta}^{k+1}$ by $\boldsymbol{\theta}^k - \alpha \sum_{m=1}^{M} Q_{\hat{b}_m^k}(\hat{\boldsymbol{g}}_m^k)$.
end for

Lemma 1. *Under Assumption 1, (11) holds if the following three inequalities are satisfied simultaneously:*

$$-\frac{\alpha}{2} + \frac{1}{2}\alpha\rho_1 + (L + 2\beta_1)(1 + \rho_2)\alpha^2 \leq 0 \tag{12a}$$

$$\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1 + {\rho_2}^{-1})\alpha^2\right]\frac{\xi_D}{\alpha^2} - \beta_D \leq 0 \tag{12b}$$

$$\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1 + {\rho_2}^{-1})\alpha^2\right]\frac{\xi_d}{\alpha^2} + \beta_{d+1} - \beta_d \leq 0, \tag{12c}$$

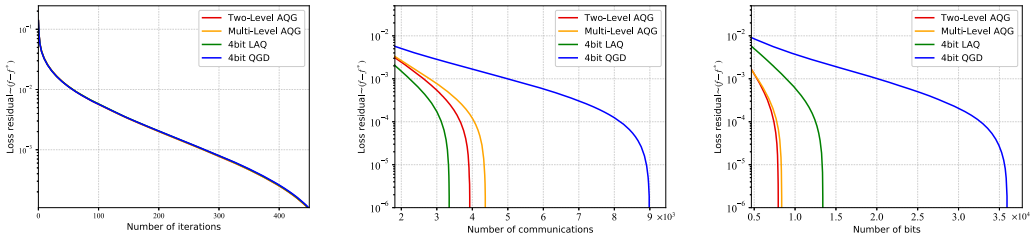
where ρ_1 and ρ_2 are constants. $\beta_d = \frac{1}{\alpha}\sum_{j=d}^{D}\xi_j, \forall d \in \{1, \ldots, D\}$. See the Appendix for proof details.

It indicates that if the stepsize α and constants $\{\xi_d\}_{d=1}^{D}$ satisfy these three inequalities, the convergence of the Lyapunov function (10) is guaranteed theoretically.

4.2 Linear Convergence With Strongly Convex Loss

The theoretical analysis under the strongly convex loss function is based on the following assumption:

Assumption 2. *Loss function $f(\boldsymbol{\theta}) = \sum_{m\in\mathbb{M}} f_m(\boldsymbol{\theta})$ is μ-strongly convex.*



(a) Loss vs iteration (b) Loss vs communication round (c) Loss vs bit

Fig. 5. Convergence of loss function with logistic regression and IID data distribution.

Under Assumption 2, there is:

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 \leq \frac{2}{\mu}[f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^*)]. \tag{13}$$

Lemma 2. *Under Assumptions 1 and 2, the following inequality holds:*

$$\begin{aligned}\mathbb{V}(\boldsymbol{\theta}^{k+1}) \leq\ & (1-c)\mathbb{V}(\boldsymbol{\theta}^k) \\ & + B\left\|\sum_{m=1}^{M} \varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + B\sum_{m\in\mathbb{M}_0^k}\left(\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_2^2 + \left\|\varepsilon_{b_{max}}\left(\boldsymbol{g}_m^k\right)\right\|_2^2\right) \\ & + B\left(\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + \sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2\right),\end{aligned} \tag{14}$$

where c and B are constants depending on μ, ρ_1, ρ_2 and parameters involved in selection criterion (5). See the Appendix for proof details.

Theorem 1. *Under Assumptions 1, 2 and Lemma 2, Lyapunov function and the quantization errors all converge at a linear rate:*

$$\left\|\varepsilon_b\left(\boldsymbol{g}_m^k\right)\right\|_\infty^2 \leq P{\tau_b}^2\sigma^k\mathbb{V}(\boldsymbol{\theta}^1) \tag{15a}$$

$$\mathbb{V}(\boldsymbol{\theta}^{k+1}) \leq \sigma^k\mathbb{V}(\boldsymbol{\theta}^1), \tag{15b}$$

where $\sigma \in (0,1)$ and τ_b is the quantization granularity with 2^b quantization levels. P is a constant based on parameters in Lemma 1. See the Appendix for proof details.

5 EXPERIMENT RESULTS

In this section, the performance of FL with the proposed AQG is evaluated with regularized logistic regression and a neural network, respectively, representing strongly convex and non-convex loss function. Experiment results demonstrate that the AQG outperforms state-of-the-art quantization algorithms in terms of reducing transmission bits and resisting client dropouts.

5.1 Experiment Setup

Datasets. In this article, we evaluate the proposed AQG method with a heterogeneous simulation dataset [7], MNIST and CIFAR10, considering both IID and non-IID data distribution. To simulate non-IID data distribution with MNIST and CIFAR10, each client is assigned only two classes of data with a balanced amount. The detailed description of the adopted dataset is provided in the Appendix.

Table 2. Performance Comparison of Gradient-Based Algorithms

Experiment setting			Iteration #	Communication #	Bit #	Transmission Reduction
Logistic Regression	IID	**Two-Level AQG**	500	3,933	**7,952**	**41%**
		Multilevel AQG	500	4,372	**8,372**	**38%**
		4-bit LAQ	500	3,354	1.34×10^4	0
		4-bit QGD	500	9,000	3.6×10^4	–*
	non-IID	**Two-Level AQG**	500	4,870	$\mathbf{1.54 \times 10^4}$	**51%**
		Multilevel AQG	500	8,273	$\mathbf{1.78 \times 10^4}$	**43%**
		4-bit LAQ	500	7,842	3.14×10^4	0
		32-bit GD[1]	500	9,000	2.88×10^5	–
Neural Network	IID	**Two-Level AQG**	2,713	854	**1,708**	**34%**
		Multilevel AQG	2,881	974	**1,928**	**25%**
		4-bit LAQ	2,784	643	2,572	0
		4-bit QGD	2,890	28,900	1.16×10^5	–
	non-IID	**Two-Level AQG**	1,319	1030	**2,060**	**44%**
		Multilevel AQG	1,702	977	**1,845**	**49%**
		4-bit LAQ	2,19	921	3,684	0
		4-bit QGD	1,251	12,510	50,040	–

[1]Since 4-bit QGD fails to converge with logistic regression and non-IID data distribution, the 32-bit vanilla GD is implemented for comparison.

*4-bit QGD definitely costs more bits compared with the baseline 4-bit LAQ.

Models. We implement logistic regression with the simulation dataset, a fully-connected network with MNIST, and a ResNet18 model with CIFAR10.

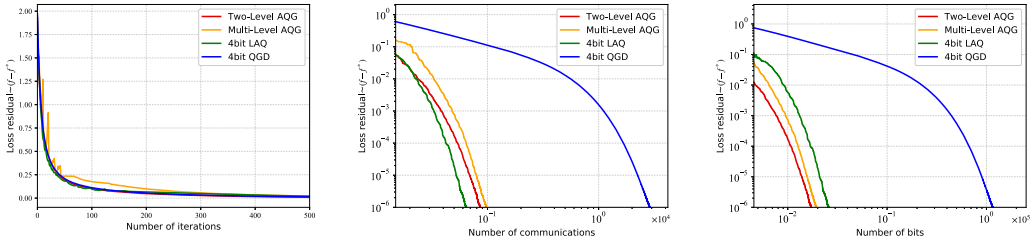
Parameters. For the AQG and LAQ, the constant parameter $D = 10$, the weights $\{\xi_d\}_{d=1}^D = 1/D$, and $M = 10$. Other standard hyperparameters of the training process are listed in Table 3 in the Appendix.

The experiment results of logistic regression and the fully connected neural network are shown in Table 2. For logistic regression, all algorithms run 500 iterations. For the fully connected network, all algorithms run 4,000 iterations, and we calculate the number of iterations, communication rounds, and transmission bits when the loss residual decreases to less than 1×10^{-6}. For both tasks, the amount of bits counted for each algorithm in Table 2 is the number of bits used to transmit **one** dimension of the uploaded gradient. Generally speaking, the proposed AQG achieves transmission reduction in all experimental settings, and the transmission reduction for non-IID data distribution is more significant than that of IID data distribution. The experiment results of ResNet18 with CIFAR10 shown in the Appendix demonstrate a similar trend.

5.2 Performance Analysis

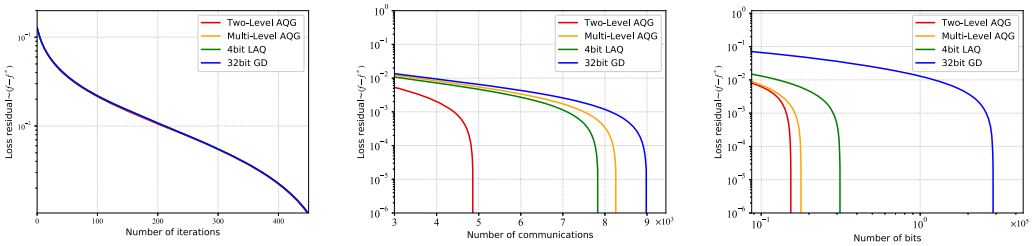
5.2.1 Performance of the AQG with IID Data Distribution. Figure 5(a) shows that with IID data distribution, the multi-level AQG and the two-level variant of AQG both reach a linear convergence rate as LAQ and QGD in strongly convex conditions. Meanwhile, AQG significantly saves transmission bits compared with 4-bit LAQ and 4-bit QGD, as shown in Figure 5(c). It can be observed from Figure 5(b) that the reduction of transmission bits is at the cost of a slight increase in communication rounds compared with LAQ, but it is worthwhile due to the significant reduction in overall transmission load.

Figure 6 shows the experiment results with IID data distribution and non-convex loss function. Similar to the results with logistic regression, the multi-level AQG and two-level AQG both require fewer bits to reach convergence without sacrificing the convergence properties of 4-bit LAQ and 4-bit QGD, as depicted in Figures 6(a) and 6(c). Meanwhile, compared with 4-bit QGD, the AQG significantly reduces communication rounds to the same order of magnitude as 4-bit LAQ, as shown in Figure 6(b).



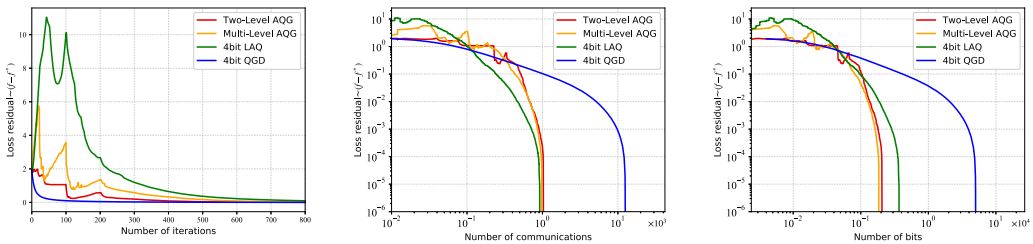
(a) Loss vs iteration (b) Loss vs communication round (c) Loss vs bit

Fig. 6. Convergence of loss function with neural network and IID data distribution.



(a) Loss vs iteration (b) Loss vs communication round (c) Loss vs bit

Fig. 7. Convergence of loss function with logistic regression and non-IID data distribution.

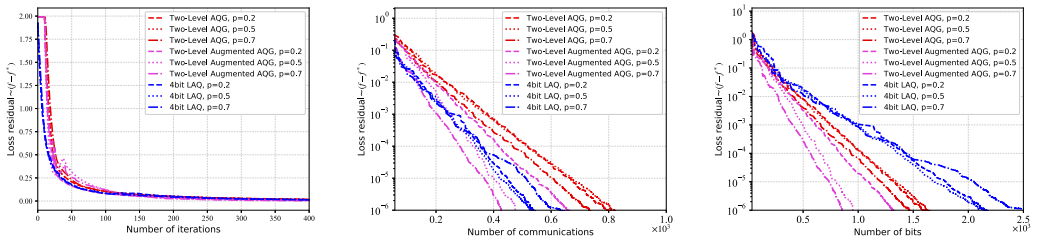


(a) Loss vs iteration (b) Loss vs communication round (c) Loss vs bit

Fig. 8. Convergence of loss function with neural network and non-IID data distribution.

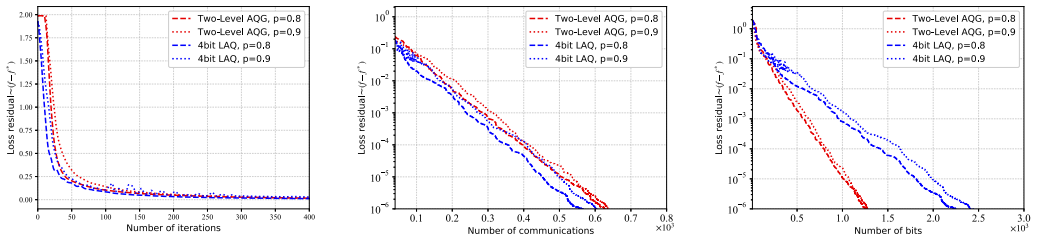
5.2.2 Performance of the AQG with non-IID Data Distribution. Figures 7 and 8 verify that the AQG works well with heterogeneous data distribution. Both variants of AQG manage to reduce the amount of transmitted bits compared with other alternatives in both strongly convex and non-convex optimization. Meanwhile, it is obvious that experiments in non-IID data distribution benefit more from the AQG compared with IID data distribution. The results are consistent with our expectation since the idea of AQGs is to utilize the inherent heterogeneity of local optimization objectives.

5.2.3 Performance of the AQG with Client Dropouts. In this subsection, we focus on the setting of wireless networks with mobile devices, in which computation and communication are both extremely expensive and client dropouts are frequent. Given these constraints, the two-level AQG is applied in experiments with client dropouts as an adaptive solution for both communication and computation efficiency. Figure 9 shows the performance of the AQG with a client dropping rate p of 0.2, 0.5, and 0.7. Experiment results demonstrate that both the AQG and Augmented AQG require



(a) Loss vs iteration (b) Loss vs communication round (c) Loss vs bit

Fig. 9. Convergence of loss function with neural network **(p = 0.2, 0.5, and 0.7)**.



(a) Loss vs iteration (b) Loss vs communication round (c) Loss vs bit

Fig. 10. Convergence of loss function with neural network **(p = 0.8 and 0.9)**.

fewer transmission bits compared with LAQ. Moreover, the Augmented AQG has a stronger ability to reduce transmission bits with the presence of such moderate client dropouts.

Figure 10 shows the performance of the AQG with a client dropping rate p of 0.8 and 0.9. Experiments show that AQG manages to achieve stable convergence with ideal rates and, at the same time, significantly reduces transmission bits even when there are only about 10% of clients participating in gradient computation at each iteration. However, we notice that the augmented version of AQG fails to converge, with a dropping rate higher than 0.8. It may be because when the dropping rate is too high, the unbiased estimation in the Augmented AQG no longer remains accurate and even induces more noise into the training. Thus, the Augmented AQG is recommended for application in FL systems in which the client dropping scale is moderate. Given the fact that the client dropping rate is not likely to be so high in most practical systems, the Augmented AQG–based method is sufficient to address the dropping problem faced by FL.

6 CONCLUSION

This article focuses on communication efficiency and the client dropout issue in FL and proposes the AQG, which not only adaptively adjusts the quantization level depending on the local gradient's update before transmission, but also appropriately amplifies transmitted gradients to limit the dropout noise. For communication efficiency, the key idea is to quantize less informative gradient with less bits and vice versa. Since the AQG fully utilizes the heterogeneity of local data distribution to reduce unnecessary transmission, it achieves a larger transmission reduction with non-IID data distribution, as expected. Compared with existing popular methods, the AQG leads to 18% to 50% of transmission reduction while keeping the desired convergence properties and shows robustness to large-scale client dropouts, with a dropping rate up to 90%. The Augmented AQG brings extra transmission reduction with moderate-scale client dropouts commonly seen in practical scenarios, which indicates the gradient amplification's effectiveness in suppressing the noise introduced by client dropouts.

Due to the aforementioned superiorities, the AQG can be used jointly with other communication-efficient methods for FL architectures, such as gradient sparsification [18, 27], client selection based on local resources [13, 23, 32] and adaptively distributing subnetworks for heterogeneous clients [6, 9]. These superiorities and flexibility indicate great potential for the proposed FL framework with the AQG. Future works include deploying the AQG jointly with such techniques in practical FL systems.

APPENDICES

A MATHEMATICAL PROOF

A.1 Proof of Lemma 1

We first derive several preliminary formulas:

(1) In the AQG, the aggregated global gradient consists of up-to-date gradients and reused gradients:

$$\begin{aligned}\sum_{m=1}^{M} Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right) &= \sum_{b=1}^{b_{max}} \sum_{m\in\mathbb{M}_b^k} Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right) + \sum_{m\in\mathbb{M}_0^k} Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\\ &= \sum_{m=1}^{M} Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right) + \sum_{b=1}^{b_{max}} \sum_{m\in\mathbb{M}_b^k} \left[Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right) - Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right]\\ &\quad + \sum_{m\in\mathbb{M}_0^k} \left[Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right) - Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right].\end{aligned} \tag{16}$$

(2) From the update rule of the AQG, there is that

$$\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k = -\alpha \sum_{m=1}^{M} Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right). \tag{17}$$

(3) The definition of the quantization error results in

$$\sum_{m=1}^{M} Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right) = \nabla f(\boldsymbol{\theta}^k) - \sum_{m=1}^{M} \varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right). \tag{18}$$

(4) With inequality $\langle \boldsymbol{a}, \boldsymbol{b}\rangle \le \frac{1}{2}\rho\|\boldsymbol{a}\|_2^2 + \frac{1}{2\rho}\|\boldsymbol{b}\|_2^2$ and (18), we have the following inequality:

$$\begin{aligned}&-\alpha\left\langle \nabla f(\boldsymbol{\theta}^k), \sum_{m=1}^{M} Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\rangle\\ &= -\alpha\left\langle \nabla f(\boldsymbol{\theta}^k), \nabla f(\boldsymbol{\theta}^k) - \sum_{m=1}^{M} \varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\rangle\\ &= -\alpha\|\nabla f(\boldsymbol{\theta}^k)\|_2^2 + \alpha\left\langle \nabla f(\boldsymbol{\theta}^k), \sum_{m=1}^{M} \varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\rangle\\ &\le -\alpha\|\nabla f(\boldsymbol{\theta}^k)\|_2^2 + \frac{\alpha\rho_1}{2}\|\nabla f(\boldsymbol{\theta}^k)\|_2^2 + \frac{\alpha}{2\rho_1}\left\|\sum_{m=1}^{M} \varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2.\end{aligned} \tag{19}$$

(5) Under Assumption 1, there is

$$
\begin{aligned}
&f(\boldsymbol{\theta}^{k+1}) - f(\boldsymbol{\theta}^k) \\
&\le \left\langle \nabla f(\boldsymbol{\theta}^k), \boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k \right\rangle + \frac{L}{2}\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\|_2^2 \\
&= \left\langle \nabla f(\boldsymbol{\theta}^k), -\alpha \sum_{m=1}^{M} Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right) \right\rangle + \frac{L}{2}\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\|_2^2 \\
&= \left\langle \nabla f(\boldsymbol{\theta}^k), -\alpha \sum_{m=1}^{M} Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right) \right\rangle + \frac{L}{2}\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\|_2^2
\end{aligned}
$$

$$
\begin{aligned}
&+ \left\langle \nabla f(\boldsymbol{\theta}^k), -\alpha \left\{ \sum_{b=1}^{b_{max}} \sum_{m \in \mathbb{M}_b^k} \left[Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right) - Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right) \right] + \sum_{m \in \mathbb{M}_0^k} \left[Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right) - Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right) \right] \right\} \right\rangle \\
&\le \left\langle \nabla f(\boldsymbol{\theta}^k), -\alpha \sum_{m=1}^{M} Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right) \right\rangle + \frac{L}{2}\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\|_2^2 + \frac{\alpha}{2}\|\nabla f(\boldsymbol{\theta}^k)\|_2^2 \\
&+ \frac{\alpha}{2} \left\| \sum_{b=1}^{b_{max}} \sum_{m \in \mathbb{M}_b^k} \left[Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right) - Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right) \right] + \sum_{m \in \mathbb{M}_0^k} \left[Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right) - Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right) \right] \right\|_2^2 .
\end{aligned} \tag{20}
$$

Then, given the following Lyapunov function of AQG:

$$
\mathbb{V}(\boldsymbol{\theta}^k) = f(\boldsymbol{\theta}^k) - f(\boldsymbol{\theta}^*) + \sum_{d=1}^{D} \sum_{j=d}^{D} \frac{\xi_j}{\alpha} \|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2, \tag{21}
$$

if we set $\beta_d = \frac{1}{\alpha} \sum_{j=d}^{D} \xi_j, \forall d \in \{1, \ldots, D\}$, then:

$$
\mathbb{V}(\boldsymbol{\theta}^k) = f(\boldsymbol{\theta}^k) - f(\boldsymbol{\theta}^*) + \sum_{d=1}^{D} \beta_d \|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2. \tag{22}
$$

Therefore, the Lyapunov function results in the following inequality:

$$
\begin{aligned}
&\mathbb{V}(\boldsymbol{\theta}^{k+1}) - \mathbb{V}(\boldsymbol{\theta}^k) \\
&= f(\boldsymbol{\theta}^{k+1}) - f(\boldsymbol{\theta}^k) \\
&\quad + \sum_{d=1}^{D} \beta_d \|\boldsymbol{\theta}^{k+1-(d-1)} - \boldsymbol{\theta}^{k-(d-1)}\|_2^2 - \sum_{d=1}^{D} \beta_d \|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2 \\
&= f(\boldsymbol{\theta}^{k+1}) - f(\boldsymbol{\theta}^k) + \beta_1 \|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\|_2^2 \\
&\quad + \sum_{d=1}^{D-1} (\beta_{d+1} - \beta_d) \|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2 - \beta_D \|\boldsymbol{\theta}^{k+1-D} - \boldsymbol{\theta}^{k-D}\|_2^2 \\
&\le -\alpha \left\langle \nabla f(\boldsymbol{\theta}^k), \sum_{m=1}^{M} Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right) \right\rangle + \frac{\alpha}{2}\|\nabla f(\boldsymbol{\theta}^k)\|_2^2 \\
&\quad + \frac{\alpha}{2} \left\| \sum_{b=1}^{b_{max}} \sum_{m \in \mathbb{M}_b^k} \left[Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right) - Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right) \right] + \sum_{m \in \mathbb{M}_0^k} \left[Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right) - Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right) \right] \right\|_2^2
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{d=1}^{D-1}(\beta_{d+1}-\beta_d)\|\boldsymbol{\theta}^{k+1-d}-\boldsymbol{\theta}^{k-d}\|_2^2-\beta_D\|\boldsymbol{\theta}^{k+1-D}-\boldsymbol{\theta}^{k-D}\|_2^2\\
&+\left(\frac{L}{2}+\beta_1\right)\|\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^{k}\|_2^2\\
=&-\alpha\left\langle\nabla f(\boldsymbol{\theta}^k),\sum_{m=1}^{M}Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\rangle+\frac{\alpha}{2}\|\nabla f(\boldsymbol{\theta}^k)\|_2^2\\
&+\sum_{d=1}^{D-1}(\beta_{d+1}-\beta_d)\|\boldsymbol{\theta}^{k+1-d}-\boldsymbol{\theta}^{k-d}\|_2^2-\beta_D\|\boldsymbol{\theta}^{k+1-D}-\boldsymbol{\theta}^{k-D}\|_2^2\\
&+\frac{\alpha}{2}\left\|\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left[Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)-Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right]+\sum_{m\in\mathbb{M}_0^k}\left[Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)-Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right]\right\|_2^2+A_1, \quad (23)
\end{aligned}
$$

where

$$
A_1=\left(\frac{L}{2}+\beta_1\right)\left\|\alpha\left\{\sum_{m=1}^{M}Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)+\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left[Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)-Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right]+\sum_{m\in\mathbb{M}_0^k}\left[Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)-Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right]\right\}\right\|_2^2
$$

From Young's Equality $\|\mathbf{a}+\mathbf{b}\|_2^2\le(1+\rho)\|\mathbf{a}\|_2^2+(1+\rho^{-1})\|\mathbf{b}\|_2^2$, there is that

$$
\begin{aligned}
A_1\le&\left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\left\|\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left[Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)-Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right]+\sum_{m\in\mathbb{M}_0^k}\left[Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)-Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right]\right\|_2^2\\
&+\left(\frac{L}{2}+\beta_1\right)(1+\rho_2)\alpha^2\left\|\sum_{m=1}^{M}Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2. \quad (24)
\end{aligned}
$$

From $\|\sum_{i=1}^n\boldsymbol{a}_i\|_2^2\le n\sum_{i=1}^n\|\boldsymbol{a}_i\|_2^2$, there is that

$$
\begin{aligned}
&\left\|\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left[Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)-Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right]+\sum_{m\in\mathbb{M}_0^k}\left[Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)-Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right]\right\|_2^2\\
&\le M\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)-Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2+M\sum_{m\in\mathbb{M}_0^k}\left\|Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)-Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2\\
&=2M\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2+2M\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2+M\sum_{m\in\mathbb{M}_0^k}\left\|Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)-Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2. \quad (25)
\end{aligned}
$$

Therefore, with (24) and (25):

$$
\begin{aligned}
\mathbb{V}(\boldsymbol{\theta}^{k+1})-\mathbb{V}(\boldsymbol{\theta}^k)\le&-\alpha\left\langle\nabla f(\boldsymbol{\theta}^k),\sum_{m=1}^{M}Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\rangle+\frac{\alpha}{2}\|\nabla f(\boldsymbol{\theta}^k)\|_2^2\\
&+\left(\frac{L}{2}+\beta_1\right)(1+\rho_2)\alpha^2\left\|\sum_{m=1}^{M}Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2+\sum_{d=1}^{D-1}(\beta_{d+1}-\beta_d)\|\boldsymbol{\theta}^{k+1-d}-\boldsymbol{\theta}^{k-d}\|_2^2-\beta_D\|\boldsymbol{\theta}^{k+1-D}-\boldsymbol{\theta}^{k-D}\|_2^2\\
&+\left[\frac{\alpha}{2}+\left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\right]\left\|\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left[Q_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)-Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right]+\sum_{m\in\mathbb{M}_0^k}\left[Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)-Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right]\right\|_2^2
\end{aligned}
$$

$$\leq -\alpha\left\langle\nabla f(\boldsymbol{\theta}^k), \sum_{m=1}^{M} Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\rangle + \frac{\alpha}{2}\|\nabla f(\boldsymbol{\theta}^k)\|_2^2 + \left(\frac{L}{2}+\beta_1\right)(1+\rho_2)\alpha^2\left\|\sum_{m=1}^{M} Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2$$
$$+ \sum_{d=1}^{D-1}(\beta_{d+1}-\beta_d)\|\boldsymbol{\theta}^{k+1-d}-\boldsymbol{\theta}^{k-d}\|_2^2 - \beta_D\|\boldsymbol{\theta}^{k+1-D}-\boldsymbol{\theta}^{k-D}\|_2^2$$
$$+ 2\left[\frac{\alpha}{2}+\left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\right]M\left(\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + \sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2\right)$$
$$+ \left[\frac{\alpha}{2}+\left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\right]M\sum_{m\in\mathbb{M}_0^k}\left\|Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)-Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2. \tag{26}$$

With the precision selection criterion (5), we have that

$$M\sum_{m\in\mathbb{M}_0^k}\left\|Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)-Q_{b_{max}}\left(\boldsymbol{g}_m^k\right)\right\|_2^2$$
$$\leq \frac{M^2}{\alpha^2M^2}\sum_{d=1}^{D}\xi_d\|\boldsymbol{\theta}^{k+1-d}-\boldsymbol{\theta}^{k-d}\|_2^2 + 3M\sum_{m\in\mathbb{M}_0^k}\left(\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\|_2^2 + \|\varepsilon_{b_{max}}(\boldsymbol{g}_m^k)\|_2^2\right), \tag{27}$$

then,

$$\mathbb{V}(\boldsymbol{\theta}^{k+1})-\mathbb{V}(\boldsymbol{\theta}^k)$$
$$\leq -\alpha\left\langle\nabla f(\boldsymbol{\theta}^k), \sum_{m=1}^{M} Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\rangle + \frac{\alpha}{2}\|\nabla f(\boldsymbol{\theta}^k)\|_2^2 + \left(\frac{L}{2}+\beta_1\right)(1+\rho_2)\alpha^2\left\|\sum_{m=1}^{M} Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2$$
$$+ \sum_{d=1}^{D-1}(\beta_{d+1}-\beta_d)\|\boldsymbol{\theta}^{k+1-d}-\boldsymbol{\theta}^{k-d}\|_2^2 - \beta_D\|\boldsymbol{\theta}^{k+1-D}-\boldsymbol{\theta}^{k-D}\|_2^2$$
$$+ 2\left[\frac{\alpha}{2}+\left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\right]M\left(\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + \sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2\right)$$
$$+ \left[\frac{\alpha}{2}+\left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\right]\frac{1}{\alpha^2}\sum_{d=1}^{D}\xi_d\|\boldsymbol{\theta}^{k+1-d}-\boldsymbol{\theta}^{k-d}\|_2^2$$
$$+ 3\left[\frac{\alpha}{2}+\left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\right]M\sum_{m\in\mathbb{M}_0^k}\left(\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_2^2 + \left\|\varepsilon_{b_{max}}(\boldsymbol{g}_m^k)\right\|_2^2\right) \tag{28}$$
$$\leq \left(-\frac{\alpha}{2}+\frac{\alpha\rho_1}{2}\right)\|\nabla f(\boldsymbol{\theta}^k)\|_2^2 + \frac{\alpha}{2\rho_1}\left\|\sum_{m=1}^{M}\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + \left(\frac{L}{2}+\beta_1\right)(1+\rho_2)\alpha^2\left\|\sum_{m=1}^{M} Q_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2$$
$$+ \sum_{d=1}^{D-1}(\beta_{d+1}-\beta_d)\|\boldsymbol{\theta}^{k+1-d}-\boldsymbol{\theta}^{k-d}\|_2^2 - \beta_D\|\boldsymbol{\theta}^{k+1-D}-\boldsymbol{\theta}^{k-D}\|_2^2$$
$$+ \left[\frac{\alpha}{2}+\left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\right]\frac{1}{\alpha^2}\sum_{d=1}^{D}\xi_d\|\boldsymbol{\theta}^{k+1-d}-\boldsymbol{\theta}^{k-d}\|_2^2$$
$$+ 3\left[\frac{\alpha}{2}+\left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\right]M\sum_{m\in\mathbb{M}_0^k}\left(\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_2^2 + \left\|\varepsilon_{b_{max}}\left(\boldsymbol{g}_m^k\right)\right\|_2^2\right)$$

$$+ 2\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1+\rho_2^{-1})\alpha^2\right] M\left(\sum_{b=1}^{b_{max}} \sum_{m\in\mathbb{M}_b^k} \left\|\varepsilon_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + \sum_{b=1}^{b_{max}} \sum_{m\in\mathbb{M}_b^k} \left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2\right) \tag{29}$$

$$\begin{aligned}
&= \left(-\frac{\alpha}{2} + \frac{\alpha\rho_1}{2}\right)\|\nabla f(\boldsymbol{\theta}^k)\|_2^2 + \frac{\alpha}{2\rho_1}\left\|\sum_{m=1}^{M} \varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + \left(\frac{L}{2} + \beta_1\right)(1+\rho_2)\alpha^2\left\|\nabla f(\boldsymbol{\theta}^k) - \sum_{m=1}^{M} \varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 \\
&\quad + \sum_{d=1}^{D-1}(\beta_{d+1} - \beta_d)\|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2 - \beta_D\|\boldsymbol{\theta}^{k+1-D} - \boldsymbol{\theta}^{k-D}\|_2^2 \\
&\quad + \left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1+\rho_2^{-1})\alpha^2\right]\frac{1}{\alpha^2}\sum_{d=1}^{D}\xi_d\|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2 \\
&\quad + 3\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1+\rho_2^{-1})\alpha^2\right] M \sum_{m\in\mathbb{M}_0^k}\left(\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_2^2 + \left\|\varepsilon_{b_{max}}(\boldsymbol{g}_m^k)\right\|_2^2\right) \\
&\quad + 2\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1+\rho_2^{-1})\alpha^2\right] M\left(\sum_{b=1}^{b_{max}} \sum_{m\in\mathbb{M}_b^k} \left\|\varepsilon_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + \sum_{b=1}^{b_{max}} \sum_{m\in\mathbb{M}_b^k} \left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2\right)
\end{aligned} \tag{30}$$

$$\begin{aligned}
&\leq \left[-\frac{\alpha}{2} + \frac{\alpha\rho_1}{2} + (L+2\beta_1)(1+\rho_2)\alpha^2\right]\|\nabla f(\boldsymbol{\theta}^k)\|_2^2 + \left[\frac{\alpha}{2\rho_1} + (L+2\beta_1)(1+\rho_2)\alpha^2\right]\left\|\sum_{m=1}^{M} \varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 \\
&\quad + \left\{\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1+\rho_2^{-1})\alpha^2\right]\frac{1}{\alpha^2}\xi_D - \beta D\right\}\|\boldsymbol{\theta}^{k+1-D} - \boldsymbol{\theta}^{k-D}\|_2^2 \\
&\quad + \sum_{d=1}^{D-1}\left\{\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1+\rho_2^{-1})\alpha^2\right]\frac{1}{\alpha^2}\xi_d + \beta_{d+1} - \beta d\right\}\|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2 \\
&\quad + 3\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1+\rho_2^{-1})\alpha^2\right] M \sum_{m\in\mathbb{M}_0^k}\left(\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_2^2 + \left\|\varepsilon_{b_{max}}(\boldsymbol{g}_m^k)\right\|_2^2\right) \\
&\quad + 2\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1+\rho_2^{-1})\alpha^2\right] M\left(\sum_{b=1}^{b_{max}} \sum_{m\in\mathbb{M}_b^k} \left\|\varepsilon_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + \sum_{b=1}^{b_{max}} \sum_{m\in\mathbb{M}_b^k} \left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2\right).
\end{aligned} \tag{31}$$

Ignoring the quantization errors, the following three inequalities should hold simultaneously for $\forall d \in \{1, \ldots, D\}$ in order to ensure that $\mathbb{V}(\boldsymbol{\theta}^{k+1}) - \mathbb{V}(\boldsymbol{\theta}^k) \leq 0$:

$$-\frac{\alpha}{2} + \frac{1}{2}\alpha\rho_1 + (L+2\beta_1)(1+\rho_2)\alpha^2 \leq 0 \tag{32a}$$

$$\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1+\rho_2{}^{-1})\alpha^2\right]\frac{\xi_D}{\alpha^2} - \beta_D \leq 0 \tag{32b}$$

$$\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1+\rho_2{}^{-1})\alpha^2\right]\frac{\xi_d}{\alpha^2} + \beta_{d+1} - \beta_d \leq 0 \tag{32c}$$

(32) provides the choice of range in terms of stepsize α and weights $\{\xi_d\}_{d=1}^D$:

$$\sum_{d=1}^{D}\xi_d \leq \min\left\{\frac{1-\rho_1}{4(1+\rho_2)}, \frac{1}{2(1+\rho_2^{-1})}\right\} \tag{33a}$$

$$\alpha \leq \min\left\{\frac{2}{L}\left[\frac{1-\rho_1}{4(1+\rho_2)} - \sum_{d=1}^{D}\xi_d\right], \frac{2}{L}\left[\frac{1}{2(1+\rho_2^{-1})} - \sum_{d=1}^{D}\xi_d\right]\right\} \tag{33b}$$

This analysis indicates that there is no need to modify these two parameters involved in LAQ [15].

A.2 Proof of Lemma 2

Under Assumption 2:

$$
\begin{aligned}
&\mathbb{V}(\boldsymbol{\theta}^{k+1}) - \mathbb{V}(\boldsymbol{\theta}^k) \\
&\le 2\mu\left[-\frac{\alpha}{2} + \frac{\alpha\rho_1}{2} + (L+2\beta_1)(1+\rho_2)\alpha^2\right]\left[f(\boldsymbol{\theta}^k) - f(\boldsymbol{\theta}^*)\right] \\
&\quad + \left[\frac{\alpha}{2\rho_1} + (L+2\beta_1)(1+\rho_2)\alpha^2\right]\left\|\sum_{m=1}^{M}\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 \\
&\quad + \beta_D\left\{\left[\frac{\alpha}{2} + \left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\right]\frac{\xi_D}{\alpha^2\beta_D} - 1\right\}\|\boldsymbol{\theta}^{k+1-D} - \boldsymbol{\theta}^{k-D}\|_2^2 \\
&\quad + \sum_{d=1}^{D-1}\beta_d\left\{\left[\frac{\alpha}{2} + \left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\right]\frac{\xi_d}{\alpha^2\beta_d} + \frac{\beta_{d+1}}{\beta_d} - 1\right\}\|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2 \\
&\quad + 3\left[\frac{\alpha}{2} + \left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\right]M\sum_{m\in\mathbb{M}_0^k}\left(\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_2^2 + \left\|\varepsilon_{b_{max}}\left(\boldsymbol{g}_m^k\right)\right\|_2^2\right) \\
&\quad + 2\left[\frac{\alpha}{2} + \left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\right]M\left(\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + \sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2\right). \quad (34)
\end{aligned}
$$

Let c and B be defined as

$$
\begin{aligned}
c = \min_{d=1,\ldots,D}\Bigg\{&2\mu\left[\frac{\alpha}{2} - \frac{\alpha\rho_1}{2} - (L+2\beta_1)(1+\rho_2)\alpha^2\right], 1 - \left[\frac{\alpha}{2} + \left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\right]\frac{\xi_D}{\alpha^2\beta_D}, \\
&1 - \left[\frac{\alpha}{2} + \left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\right]\frac{\xi_d}{\alpha^2\beta_d} + \frac{\beta_{d+1}}{\beta_d}\Bigg\}. \quad (35a)
\end{aligned}
$$

$$
B = \max\left\{\frac{\alpha}{2\rho_1} + (L+2\beta_1)(1+\rho_2)\alpha^2, 3M\left[\frac{\alpha}{2} + \left(\frac{L}{2}+\beta_1\right)(1+\rho_2^{-1})\alpha^2\right]\right\}. \quad (35b)
$$

Then:

$$
\begin{aligned}
\mathbb{V}(\boldsymbol{\theta}^{k+1}) - \mathbb{V}(\boldsymbol{\theta}^k) \le &-c\left[f(\boldsymbol{\theta}^k) - f(\boldsymbol{\theta}^*) + \sum_{d=1}^{D}\beta_d\|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2\right] \\
&+ B\left\|\sum_{m=1}^{M}\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + B\sum_{m\in\mathbb{M}_0^k}\left(\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_2^2 + \left\|\varepsilon_{b_{max}}\left(\boldsymbol{g}_m^k\right)\right\|_2^2\right) \\
&+ B\left(\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + \sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2\right) \quad (36)
\end{aligned}
$$

$$
\begin{aligned}
&= -c\,\mathbb{V}(\boldsymbol{\theta}^k) \\
&\quad + B\left\|\sum_{m=1}^{M}\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + B\sum_{m\in\mathbb{M}_0^k}\left(\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_2^2 + \left\|\varepsilon_{b_{max}}\left(\boldsymbol{g}_m^k\right)\right\|_2^2\right) \\
&\quad + B\left(\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + \sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2\right).
\end{aligned} \tag{37}
$$

Thus,

$$
\begin{aligned}
\mathbb{V}(\boldsymbol{\theta}^{k+1}) &\le (1-c)\mathbb{V}(\boldsymbol{\theta}^k) \\
&\quad + B\left\|\sum_{m=1}^{M}\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + B\sum_{m\in\mathbb{M}_0^k}\left(\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_2^2 + \left\|\varepsilon_{b_{max}}\left(\boldsymbol{g}_m^k\right)\right\|_2^2\right) \\
&\quad + B\left(\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + \sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2\right).
\end{aligned} \tag{38}
$$

A.3 Proof of Theorem 1

This part proves that (15) holds for any $k \ge 0$ if the following inequalities are satisfied:

$$
4BMP\tau_{b_{max}}^2 + BMP\sum_{b=1}^{b_{max}}\tau_{b_m^k}^2 \le \sigma_2 - \sigma_1. \tag{39a}
$$

$$
\frac{24L^2}{\mu} + 18\tau_{b_{max}-b_m^k}^2 + 3\tau_{b_{max}}^2 \le \sigma_2. \tag{39b}
$$

$$
\alpha \ge \frac{\mu}{4L^2M^2}. \tag{39c}
$$

It is assumed that for any $k \ge 1$, (15) holds for $k-1$. Let $\sigma_1 = 1-c$; there is that

$$
\begin{aligned}
&\mathbb{V}(\boldsymbol{\theta}^{k+1}) \le \sigma_1\mathbb{V}(\boldsymbol{\theta}^k) \\
&\quad + B\left\|\sum_{m=1}^{M}\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + B\sum_{m\in\mathbb{M}_0^k}\left(\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_2^2 + \left\|\varepsilon_{b_{max}}\left(\boldsymbol{g}_m^k\right)\right\|_2^2\right) \\
&\quad + B\left(\sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{\hat{b}_m^k}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2 + \sum_{b=1}^{b_{max}}\sum_{m\in\mathbb{M}_b^k}\left\|\varepsilon_{b_{max}}\left(\hat{\boldsymbol{g}}_m^k\right)\right\|_2^2\right)
\end{aligned} \tag{40}
$$

$$
\begin{aligned}
&\le \sigma_1\sigma_2^{k-1}\mathbb{V}(\boldsymbol{\theta}^1) + 4BMP\tau_{b_{max}}^2\sigma_2^{k-1}\mathbb{V}(\boldsymbol{\theta}^1) + BMP\sum_{b=1}^{b_{max}}\tau_{b_m^k}^2\sigma_2^{k-1}\mathbb{V}(\boldsymbol{\theta}^1) \\
&= \left(\sigma_1 + 4BMP\tau_{b_{max}}^2 + BMP\sum_{b=1}^{b_{max}}\tau_{b_m^k}^2\right)\sigma_2^{k-1}\mathbb{V}(\boldsymbol{\theta}^1) \le \sigma_2^k\mathbb{V}(\boldsymbol{\theta}^1),
\end{aligned} \tag{41}
$$

where $\sigma_2 \ge \sigma_1 + 4BMP\tau_{b_{max}}^2 + BMP\sum_{b=1}^{b_{max}}\tau_{b_m^k}^2$.

Under Assumptions 1 and 2, the following inequality holds for any $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ because of convexity:

$$\begin{aligned}\|\nabla f_m(\boldsymbol{\theta}_1) - \nabla f_m(\boldsymbol{\theta}_2)\|_\infty &\leq \left\|\sum_{m=1}^{M}(\nabla f_m(\boldsymbol{\theta}_1) - \nabla f_m(\boldsymbol{\theta}_2))\right\|_\infty \\ &= \|\nabla f(\boldsymbol{\theta}_1) - \nabla f(\boldsymbol{\theta}_2)\|_\infty \\ &\leq L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_\infty,\ \forall m \in \{1,\ldots,M\}. \end{aligned} \tag{42}$$

With (42) and the proposed precision selection criterion (5), there is that

$$\begin{aligned}&\left\|\nabla f_m(\boldsymbol{\theta}^{k+1}) - Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_\infty \\ &= \left\|\nabla f_m(\boldsymbol{\theta}^{k+1}) - f_m(\boldsymbol{\theta}^k) + f_m(\boldsymbol{\theta}^k) - Q_{b_{max}}(\boldsymbol{g}_m^k) + Q_{b_{max}}(\boldsymbol{g}_m^k) - Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_\infty \\ &\leq \left\|\nabla f_m(\boldsymbol{\theta}^{k+1}) - f_m(\boldsymbol{\theta}^k)\right\|_\infty + \left\|f_m(\boldsymbol{\theta}^k) - Q_{b_{max}}\left(\boldsymbol{g}_m^k\right)\right\|_\infty + \left\|Q_{b_{max}}\left(\boldsymbol{g}_m^k\right) - Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_\infty \\ &\leq L\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\|_\infty + \left\|\varepsilon_{b_{max}}\left(\boldsymbol{g}_m^k\right)\right\|_\infty \\ &\quad + \sqrt{\frac{1}{\alpha^2 M^2}\sum_{d=1}^{D}\xi_d\|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2 + 3\left(\left\|\varepsilon_{b_{max}-b_m^k}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_2^2 + \left\|\varepsilon_{b_{max}-b_m^k}(\boldsymbol{g}_m^k)\right\|_2^2\right)} \qquad (43) \\ &\leq L\sqrt{\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^* + \boldsymbol{\theta}^* - \boldsymbol{\theta}^k\|_2^2} + \left\|\varepsilon_{b_{max}}\left(\boldsymbol{g}_m^k\right)\right\|_\infty \\ &\quad + \sqrt{\frac{1}{\alpha^2 M^2}\sum_{d=1}^{D}\xi_d\|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2 + 3\left(\left\|\varepsilon_{b_{max}-b_m^k}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_2^2 + \left\|\varepsilon_{b_{max}-b_m^k}(\boldsymbol{g}_m^k)\right\|_2^2\right)} \qquad (44) \\ &\leq L\sqrt{2\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^*\|_2^2 + 2\|\boldsymbol{\theta}^* - \boldsymbol{\theta}^k\|_2^2} + \left\|\varepsilon_{b_{max}}\left(\boldsymbol{g}_m^k\right)\right\|_\infty \\ &\quad + \sqrt{\frac{1}{\alpha^2 M^2}\sum_{d=1}^{D}\xi_d\|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2 + 3\left(\left\|\varepsilon_{b_{max}-b_m^k}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_\infty^2 + \left\|\varepsilon_{b_{max}-b_m^k}(\boldsymbol{g}_m^k)\right\|_\infty^2\right)}. \qquad (45)\end{aligned}$$

Under Assumption 2 with (13),

$$\begin{aligned}&\left\|\nabla f_m(\boldsymbol{\theta}^{k+1}) - Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_\infty^2 \\ &\leq \frac{12L^2}{\mu}\left[f(\boldsymbol{\theta}^{k+1}) - f(\boldsymbol{\theta}^*) + f(\boldsymbol{\theta}^k) - f(\boldsymbol{\theta}^*)\right] + 3\left\|\varepsilon_{b_{max}}\left(\boldsymbol{g}_m^k\right)\right\|_\infty^2 \\ &\quad + \frac{3}{\alpha^2 M^2}\sum_{d=1}^{D}\xi_d\|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2 + 9\left(\left\|\varepsilon_{b_{max}-b_m^k}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_\infty^2 + \left\|\varepsilon_{b_{max}-b_m^k}(\boldsymbol{g}_m^k)\right\|_\infty^2\right) \qquad (46) \\ &\leq \frac{12L^2}{\mu}\left[f(\boldsymbol{\theta}^{k+1}) - f(\boldsymbol{\theta}^*) + f(\boldsymbol{\theta}^k) - f(\boldsymbol{\theta}^*) + \frac{\mu}{4L^2\alpha^2M^2}\sum_{d=1}^{D}\xi_d\|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2\right] \\ &\quad + 18P\tau_{b_{max}-b_m^k}^2\sigma_2^{k-1}\mathbb{V}(\boldsymbol{\theta}^1) + 3P\tau_{b_{max}}^2\sigma_2^{k-1}\mathbb{V}(\boldsymbol{\theta}^1). \qquad (47)\end{aligned}$$

With $\alpha \geq \frac{\mu}{4L^2M^2}$, $\frac{\mu\xi_d}{4L^2\alpha^2M^2} \leq \frac{\xi_d}{\alpha} \leq \sum_{j=d}^{D}\frac{\xi_j}{\alpha}$:

$$\left\|\nabla f_m(\boldsymbol{\theta}^{k+1}) - Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_\infty^2$$

$$\leq \frac{12L^2}{\mu}\left[f(\boldsymbol{\theta}^{k+1}) - f(\boldsymbol{\theta}^*) + f(\boldsymbol{\theta}^k) - f(\boldsymbol{\theta}^*) + \sum_{d=1}^{D}\sum_{j=d}^{D}\frac{\xi_j}{\alpha}\|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2\right]$$
$$+ 18P\tau^2_{b_{max}-b_m^k}\sigma_2^{k-1}\mathbb{V}(\boldsymbol{\theta}^1) + 3P\tau^2_{b_{max}}\sigma_2^{k-1}\mathbb{V}(\boldsymbol{\theta}^1)$$
$$\leq \frac{12L^2}{\mu}\left[\mathbb{V}(\boldsymbol{\theta}^{k+1}) + \mathbb{V}(\boldsymbol{\theta}^k)\right] + 18P\tau^2_{b_{max}-b_m^k}\sigma_2^{k-1}\mathbb{V}(\boldsymbol{\theta}^1) + 3P\tau^2_{b_{max}}\sigma_2^{k-1}\mathbb{V}(\boldsymbol{\theta}^1)$$
$$\leq \frac{24L^2}{\mu}\sigma_2^{k-1}\mathbb{V}(\boldsymbol{\theta}^1) + 18P\tau^2_{b_{max}-b_m^k}\sigma_2^{k-1}\mathbb{V}(\boldsymbol{\theta}^1) + 3P\tau^2_{b_{max}}\sigma_2^{k-1}\mathbb{V}(\boldsymbol{\theta}^1)$$
$$= \left(\frac{24L^2}{\mu P} + 18\tau^2_{b_{max}-b_m^k} + 3\tau^2_{b_{max}}\right)P\sigma_2^{k-1}\mathbb{V}(\boldsymbol{\theta}^1) \leq P\sigma_2^k\mathbb{V}(\boldsymbol{\theta}^1). \tag{48}$$

Thus,

$$\left\|\varepsilon_b(\boldsymbol{g}_m^k)\right\|_\infty^2 \leq {\tau_b}^2\left\|\nabla f_m(\boldsymbol{\theta}^{k+1}) - Q_{\hat{b}_m^{k-1}}\left(\hat{\boldsymbol{g}}_m^{k-1}\right)\right\|_\infty^2 \leq P{\tau_b}^2\sigma_2^k\mathbb{V}(\boldsymbol{\theta}^1). \tag{49}$$

B SUPPLEMENTARY EXPERIMENTAL INFORMATION

B.1 Hyperparameters

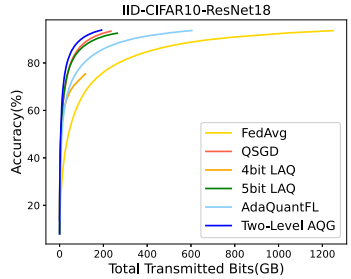
Table 3. Hyperparameters in Training Process

Dataset	MNIST		CIFAR10	
Model	FC		ResNet18	
Hidden Size	[784, 10]		[64, 128, 256, 512]	
Data Distribution	IID	non-IID	IID	non-IID
Global Epoch $\boldsymbol{E}$	4,000	4,000	5,000	100
Local Batch Size $\boldsymbol{B}$	/	/	100	10
Optimizer	GD	GD	SGD	SGD
Momentum	/	/	0.9	0.9
Weight Decay	/	/	5.00E-04	5.00E-04
Learning Rate η	0.02	0.02	0.1	0.1

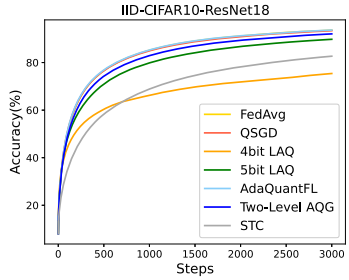
B.2 Experiment Results with IID/Non-IID CIFAR10

Note that for experiments with CIFAR10 and ResNet18, we adopt stochastic gradient descend (SGD) and we add two state-of-the-art baselines: AdaQuantFL [14] and STC [24]. The two-level AQG represents anAQG with 6/4 bit transmission for IID CIFAR10 and 5/3 bit transmission for non-IID CIFAR10.

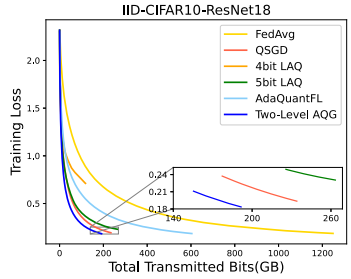
From Figures 11 and 12, we can conclude that the proposed AQG achieves a significant transmission reduction on more complex datasets and models as compared with baselines, including QSGD, fixed-bit LAQ, and AdaQuantFL. Specifically, the transmission reduction is 18.13% for IID CIFAR10 and 25.53% for non-IID CIFAR10, as shown in Figures 11(c) and 12(c), respectively. Note that since STC fails to achieve the same convergence compared with other baselines, we do not include it in the comparison for transmitted bits. The slow convergence of STC and 4-bit LAQ verifies the necessity and effectiveness of our well-designed precision selection criterion (5), which achieves fast convergence with similar low-bit transmission but without degradation of model performance.



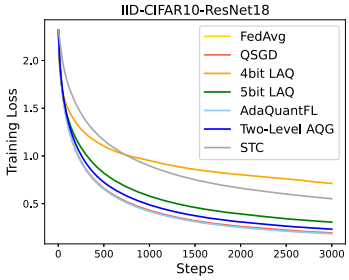
(a) Accuracy vs Total Transmitted Bits



(b) Accuracy vs Steps

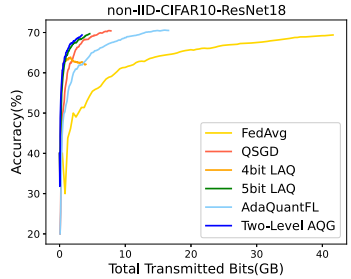


(c) Training Loss vs Total Transmitted Bits

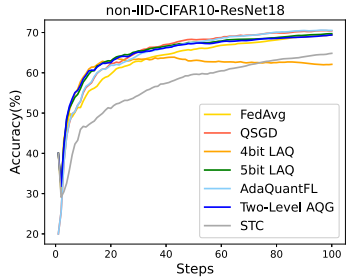


(d) Training Loss vs Steps

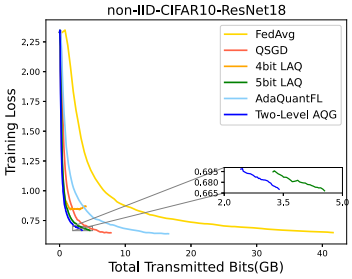
Fig. 11. Convergence of loss function with ResNet18 and IID CIFAR10.



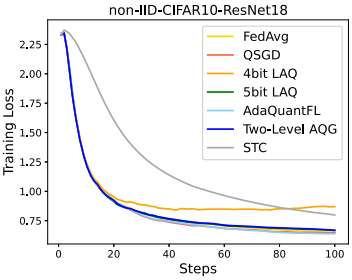
(a) Accuracy vs Total Transmitted Bits



(b) Accuracy vs Steps



(c) Training Loss vs Total Transmitted Bits



(d) Training Loss vs Steps

Fig. 12. Convergence of loss function with ResNet18 and non-IID CIFAR10.

B.3 Comparison of Converged Accuracy

Table 4 compares the proposed AQG with several baselines in terms of converged test accuracy.

Table 4. Comparison of Converged Test Accuracy

	Two-level AQG	LAQ	Q(S)GD	FedAvg
IID-MNIST-FC	90.81%	90.82%	90.79%	90.77%
non-IID-MNIST-FC	90.78%	90.77%	90.77%	90.62%
IID-CIFAR10-ResNet18	92.95%	92.53%	93.39%	93.61%
non-IID-CIFAR10-ResNet18	69.38%	69.71%	70.41%	69.39%

FC denotes the 2-layer fully connected neural network adopted in the main text. The quantization levels of two-level AQG is 4/2 bits for MNIST, 6/4 bits for IID CIFAR10 and 5/3 bits for non-IID CIFAR10.

B.4 Heterogeneous Simulation Dataset

To simulate non-IID data distribution, a heterogeneous simulation dataset is used for logistic regression. The three binary classification datasets listed in Table 5 are used together in order to simulate non-IID data distribution as Chen et al. do in the evaluation of LAQ [7]. The number of features is preprocessed to be equal to the minimal number of features among all three datasets, and each dataset is uniformly distributed across six clients.

Table 5. Heterogeneous Simulation Datasets Used for Logistic Regression

Dataset	# features	# samples	client index
Adult fat [17]	113	1605	1,2,3,4,5,6
Ionosphere [26]	34	351	7,8,9,10,11,12
Derm [11]	34	358	13,14,15,16,17,18

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. Vienna, Austria, ACM, 308–318.

[2] Alham Fikri Aji and Kenneth Heafield. 2017. Sparse communication for distributed gradient descent. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark. Association for Computational Linguistics, 440–445.

[3] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. 2017. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., Long Beach, CA, USA.

[4] Jeremy Bernstein, Yu Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. 2018. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. PMLR, Stockholmsmässan, Stockholm, Sweden, 560–569.

[5] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. Dallas, TX, ACM, 1175–1191.

[6] Nader Bouacida, Jiahui Hou, Hui Zang, and Xin Liu. 2020. Adaptive federated dropout: Improving communication efficiency and generalization for federated learning. arXiv preprint arXiv:2011.04050.

[7] Tianyi Chen, Georgios B. Giannakis, Tao Sun, and Wotao Yin. 2018. LAG: Lazily aggregated gradient for communication-efficient distributed learning. arXiv preprint arXiv:1805.09965.

[8] Kewei Cheng, Tao Fan, Yilun Jin, Yang Liu, Tianjian Chen, and Qiang Yang. 2019. Secureboost: A lossless federated learning framework. arXiv preprint arXiv:1901.08755.

[9] Enmao Diao, Jie Ding, and Vahid Tarokh. 2020. HeteroFL: Computation and communication efficient federated learning for heterogeneous clients. arXiv preprint arXiv:2010.01264.
[10] Zhaoyang Du, Celimuge Wu, Tsutomu Yoshinaga, Kok-Lim Alvin Yau, Yusheng Ji, and Jie Li. 2020. Federated learning for vehicular Internet of Things: Recent advances and open issues. *IEEE Open Journal of the Computer Society* 1 (2020), 45–61.
[11] H. A. Güvenir, G. Demiröz. and N. Ilter. 1998. Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals. *Artificial Intelligence in Medicine* 13, 3 (1998), 147–165.
[12] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. 2019. Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604.
[13] Yinghui He, Jinke Ren, Guanding Yu, and Jiantao Yuan. 2020. Resource allocation for wireless federated edge learning based on data importance. In *IEEE Global Communications Conference*. IEEE, 1–6.
[14] Divyansh Jhunjhunwala, Advait Gadhikar, Gauri Joshi, and Yonina C. Eldar. 2021. Adaptive quantization of model updates for communication-efficient federated learning. In *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, Toronto, Ontario, Canada, 3110–3114.
[15] Sun Jun, Chen Tianyi, Georgios B. Giannakis, Yang Qinmin, and Yang Zaiyue. 2022. Lazily aggregated quantized gradient innovation for communication-efficient federated learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2022), 2031–2044.
[16] Georgios A. Kaissis, Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* 2, 6 (2020), 305–311.
[17] Ron Kohavi. 1996. Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* 96 (1996), 202–207.
[18] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, Vol. 2. 429–450.
[19] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J. Dally. 2018. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *Proceedings of International Conference on Learning Representations*, Vancouver, BC, Canada.
[20] Yang Liu, Anbu Huang, Yun Luo, He Huang, Youzhi Liu, Yuanyuan Chen, Lican Feng, Tianjian Chen, Han Yu, and Qiang Yang. 2020. Fedvision: An online visual object detection platform powered by federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. AAAI Press, New York, NY, USA, 13172–13179.
[21] Yang Liu, Yan Kang, Chaoping Xing, Tianjian Chen, and Qiang Yang. 2020. A secure federated transfer learning framework. *IEEE Intelligent Systems* 35, 4 (2020), 70–82.
[22] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Vol. 54. PMLR, Fort Lauderdale, FL, 1273–1282.
[23] Takayuki Nishio and Ryo Yonetani. 2019. Client selection for federated learning with heterogeneous resources in mobile edge. In *IEEE International Conference on Communications (ICC'19)*. IEEE, 1–7.
[24] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. 2020. Robust and communication-efficient federated learning from non-IID data. *IEEE Transactions on Neural Networks and Learning Systems* 31, 9 (2020), 3400–3413.
[25] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 2014. 1-Bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Fifteenth Annual Conference of the International Speech Communication Association*. Citeseer, Singapore, 1058–1062.
[26] Vincent G. Sigillito, Simon P. Wing, Larrie V. Hutton, and Kile B. Baker. 1989. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest* 10, 3 (1989), 262–266.
[27] Navjot Singh, Deepesh Data, Jemin George, and Suhas Diggavi. 2020. SPARQ-SGD: Event-triggered and compressed communication in decentralized optimization. In *59th IEEE Conference on Decision and Control (CDC'20)*. IEEE, Jeju Island, Republic of Korea, 3449–3456.
[28] Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. 2018. Sparsified SGD with memory. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Montréal, Canada, 4452–4463.
[29] Hongyi Wang, Scott Sievert, Zachary Charles, Shengchao Liu, Stephen Wright, and Dimitris Papailiopoulos. 2018. ATOMO: Communication-efficient learning via atomic sparsification. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Montréal, Quebec, Canada, Curran Associates Inc., Red Hook, 9872–9883.
[30] Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. 2018. Gradient sparsification for communication-efficient distributed optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Montréal, Quebec, Canada, Curran Associates Inc., 1306–1316.

[31] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology* 10, 2 (2019), 1–19.

[32] Tongxin Zhu, Jianzhong Li, Zhipeng Cai, Yingshu Li, and Hong Gao. 2020. Computation scheduling for wireless powered mobile edge computing networks. In *IEEE Conference on Computer Communications*. IEEE, 596–605.

Received April 2021; revised November 2021; accepted January 2022