

Communication Efficient Federated Learning with Adaptive Quantization

YUZHU MAO, Tsinghua Shenzhen International Graduate School, Tsinghua University, China and Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University, China

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, Department of AI, WeBank, 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 and Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University, China

Federated learning (FL) has attracted tremendous attentions in recent years due to its privacy preserving measures and great potentials in some distributed but privacy-sensitive applications like finance and health. However, high communication overloads for transmitting high-dimensional networks and extra security masks remains a bottleneck of FL. This paper proposes a communication-efficient FL framework with Adaptive Quantized Gradient (AQG) which adaptively adjusts the quantization level based on local gradient's update to fully utilize the heterogeneousness of local data distribution for reducing unnecessary transmissions. Besides, the client dropout issues are taken into account and the 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 25%-50% of additional transmission reduction as compared to existing popular methods including Quantized Gradient Descent (QGD) and Lazily Aggregated Quantized (LAQ) gradient-based method without deteriorating convergence properties. Particularly, experiments with heterogeneous data distributions corroborate a more significant transmission reduction compared with independent identical data distributions. Meanwhile, 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; Networks** → Network reliability.

Additional Key Words and Phrases: federated learning, distributed learning, quantization

Authors' addresses: Yuzhu Mao, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China, 518055, Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University, Shenzhen, China, 518055, myz20@mails.tsinghua.edu.cn; Zihao Zhao, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China, 518055, Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University, Shenzhen, China, 518055, kevinzhaozh1998@gmail.com; Guangfeng Yan, City University of Hong Kong, Hong Kong, China, City University of Hong Kong Shenzhen Research Institute, Shenzhen, China, gfyang2-c@my.cityu.edu.hk; Yang Liu, Department of AI, WeBank, Shenzhen, China, yangliu@webank.com; Tian Lan, Department of Electrical and Computer Engineering, George Washington University, DC, USA, tlan@gwu.edu; Linqi Song, City University of Hong Kong, Hong Kong, China, City University of Hong Kong Shenzhen Research Institute, Shenzhen, China, linqi.song@cityu.edu.hk; Wenbo Ding*, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China, 518055, Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University, Shenzhen, China, 518055, ding.wenbo@sz.tsinghua.edu.cn.

1 INTRODUCTION

The deployment of Internet of things (IoT), ubiquitous sensing, edge computing and many other distributed systems have enabled the fast development of distributed learning techniques in recent years[10, 12, 17]. The distributed learning could fully utilize the low-cost computing resources throughout the network and achieve comparable performance with the centralized learning. Nevertheless, the leakage of the data, gradient, and even model 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, health, and etc. To this end, the federated learning (FL) which prevents privacy leakage by avoiding data exposition has been proposed by Google and other researchers, and attracted tremendous attentions from both academia and industry[19].

Many approaches like differential privacy[1], secret sharing techniques[5] and homomorphic encryption[18] have been developed to mask the transmitted gradients and can almost well address the security issues in FL. However, high-dimensional neural networks and extra security masks[8, 14, 28] may lead to high communication overhead, which becomes a main bottleneck of FL systems. In this context, the communication-efficient learning algorithms have been proposed mainly to reduce the 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 the Quantized Stochastic Gradient Descent (QSGD)[3], 1-bit SGD[21] and SignSGD[4], etc. 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 where 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[25]. Although LAQ 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 paper proposes a communication efficient FL framework with Adaptive Quantized Gradient (AQG), where the quantization level is adjusted according to the local gradient's updates adaptively. Specifically, gradients with larger amount of updates are quantized and transmitted with more bits, and vice versa. Besides, this paper 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 AQG with the presence of the noise introduced by client dropouts, the proposed FL framework with AQG is augmented by a variance-reduced method, where 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. Meanwhile, AQG is robust to a client dropping rate up to 90% empirically, and the Augmented AQG with gradient amplification does act as a competitive solution to achieve an even more significant transmission reduction with moderate clients dropping scale commonly seen in practical FL scenarios.

The remainder of the paper is organized as follows. Section 2 provides the FL system overviews and discusses our motivations. The proposed Adaptive Quantized Gradient method is elaborated in section 3. 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 conclude this paper in section 6.

Notation. The notations involved in this paper are listed in Table 1.

Table 1. Notations

\mathbf{g}_m^k	gradient computed by client m at iteratoin k
$\hat{\mathbf{g}}_m^k$	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 $\hat{\mathbf{g}}_m^k$
$Q_b(\mathbf{g}_m^k)$	\mathbf{g}_m^k quantized with b bits
θ^k	the aggregated global model broadcasted at iteration k
$\varepsilon_b(\mathbf{g}_m^k)$	quantization error ($Q_b(\mathbf{g}_m^k) - \mathbf{g}_m^k$)
\mathbb{M}	clients set
\mathbb{M}_b^k	subset of clients uploading gradients with b bits at iteration k
p	clients 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}

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 Fig. 1. Similar to most distributed learning systems, FL system uses a server to receive locally-computed gradients and update 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]. Besides, 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 Fig. 1. Thus, the robustness to client dropout is another practical requirement for FL systems[5].

2.2 Motivations

FL is bottlenecked by the high communication overheads and limited device reliability. The lack of efficient transmission and robustness to client dropouts may lead to slow, expensive and unstable learning. In this paper, the FL framework with the proposed AQG method provides opportunities for communication-efficient FL with large-scale of client dropouts.

Firstly, 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 descend at different rates. Therefore, adaptively adjusting the quantization level according to gradient's update amount provides a more efficient way to communicate with the server by quantizing slowly-varying gradients with less amount of bits.

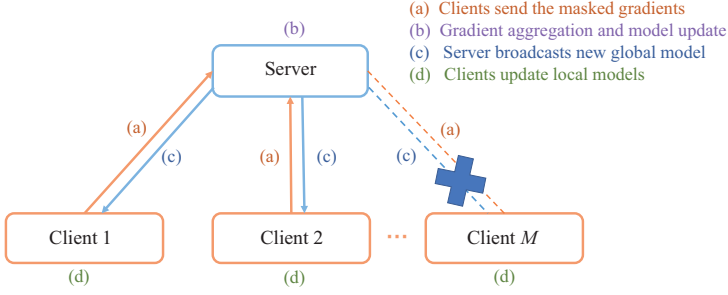


Fig. 1. Typical architecture for a FL system.

Secondly, AQG aims to address the noise induced by client dropouts. When a client dropout occurs, all coordinates of transmitted gradient are lost, which can be regarded as an extreme example of gradient sparsification[2, 16, 24, 26]. In order to limit the variance increase of a sparsified gradient, Wangni *et al.* proposed to keep the unbiasedness of the sparsified gradient by appropriately amplifying the remaining coordinates[27]. Inspired by this idea, 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 the transmission overheads, a multilevel adaptive quantization scheme is proposed in this section. As illustrated in Fig. 2, the FL system with AQG can be implemented as follows. At iteration k , the server broadcasts global model θ^k to all clients. Each client computes gradient g_m^k by taking all its local data \mathbf{X}_m as a full batch:

$$g_m^k = \nabla f_m(\mathbf{X}_m; \theta^k) \quad (1)$$

After the gradient computation, each client needs to make **two decisions**: (1) is it necessary to send its quantized gradient? (2) how many bits b_m^k should be used to quantize and send its newly-computed gradient? In particular, the first decision is the key idea in LAQ[25]. In this paper, it is considered as a special case of the second decision, where b_m^k is chosen as zero if the client decides to send nothing.

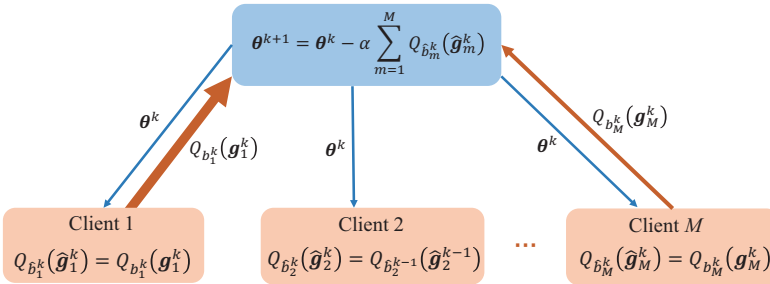


Fig. 2. FL with AQG.

If client m chooses a non-zero b_m^k and updates its newly-quantized gradient, then $Q_{b_m^k}(g_m^k)$ is one of the quantized gradients that actually participate in gradient aggregation on the server side

at iteration k . Otherwise, the server reuses the old quantized-gradient $Q_{\hat{b}_m^{k-1}}(\hat{\mathbf{g}}_m^{k-1})$ from the last iteration to represent client m in the aggregation. In summary, an iteration step of proposed AQG is as follows:

$$\text{Gradients Update} \quad Q_{\hat{b}_m^k}(\hat{\mathbf{g}}_m^k) = \begin{cases} Q_{b_m^k}(\mathbf{g}_m^k), & m \in \mathbb{M} \setminus \mathbb{M}_0^k \\ Q_{\hat{b}_m^{k-1}}(\hat{\mathbf{g}}_m^{k-1}), & m \in \mathbb{M}_0^k \end{cases} \quad (2)$$

$$\text{Gradients Aggregation} \quad \boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \alpha \sum_{m \in \mathbb{M}} Q_{\hat{b}_m^k}(\hat{\mathbf{g}}_m^k) \quad (3)$$

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 , $Q_{\hat{b}_m^k}(\hat{\mathbf{g}}_m^k)$ 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 is that:

- 1) For clients belonging to $\mathbb{M} \setminus \mathbb{M}_0^k$, the precision levels (quantization levels) of their new updates $Q_{b_m^k}(\mathbf{g}_m^k)$ are not fixed, but adaptively adjusted depending on \mathbf{g}_m^k 's innovations—the difference between the newly-quantized gradient and the last quantized gradient sent to the server. It motivates a need for not only a quantization scheme as previous work, but also a **precision selection criterion** to decide the quantization level of each newly-computed gradient;
- 2) For FL scenario where client dropouts is relatively frequent, methods to limit the noise introduced by gradients lossing are also in great need.

The following part of this section presents the precision selection criterion developed in this paper and the quantization scheme applied in the proposed AQG. At last, 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 $Q_b(\mathbf{g}_m^k)$ and the last upload $Q_b(\hat{\mathbf{g}}_m^{k-1})$, where b is the fixed number of bits after quantization[25]. In order to decide whether client m needs to upload its newly-quantized gradient $Q_b(\mathbf{g}_m^k)$ at iteration k , LAQ develops a communication selection criterion as follows:

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

where $\varepsilon_b(\hat{\mathbf{g}}_m^{k-1})$ and $\varepsilon_b(\mathbf{g}_m^k)$ 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 $Q_b(\mathbf{g}_m^k)$ at iteration k only when the difference between $Q_b(\mathbf{g}_m^k)$ and the last upload $Q_b(\hat{\mathbf{g}}_m^{k-1})$ is larger than a threshold, which takes the quantization error and global model's innovation into account[25].

This paper extends the single precision level LAQ with communication selection criterion (4) to multilevel adaptive quantization for transmitted gradients. The key idea of AQG is that under a pre-set upper bound b_{max} for the number of bits after quantization, gradients with smaller innovations can be quantized with less 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 \mathbf{g}_m^k , we develop the following precision selection criterion:

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

As illustrated in Fig. 3, the proposed precision selection criterion (5) works in the following ways:

- 1) For any $\bar{b} \in [1, \dots, b_{\max}-1]$, satisfying (5) with $b = \bar{b}+1$ will necessarily satisfy (5) with $b = \bar{b}$, but not vice versa. The reason is that for a given \mathbf{g}_m^k , there is always $\varepsilon_{b_{\max}-(\bar{b}+1)+1}(\mathbf{g}_m^k) = \varepsilon_{b_{\max}-\bar{b}}(\mathbf{g}_m^k) \geq \varepsilon_{b_{\max}-\bar{b}+1}(\mathbf{g}_m^k)$ due to more error brought by more aggressive quantization.
- 2) Precision selection criterion (5) with $b = 1$ acts as communication selection criterion in AQG. Specifically, if (5) with $b = 1$ does not hold for client m , then its gradient update at iteration k is skipped.

Therefore, client subsets devided by the proposed precision criterion form the client set \mathbb{M} without overlaps:

$$\mathbb{M}_0^k \cup \mathbb{M}_1^k \cup \mathbb{M}_2^k \cup \dots \cup \mathbb{M}_{b_{\max}}^k = \mathbb{M} \quad (6)$$

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

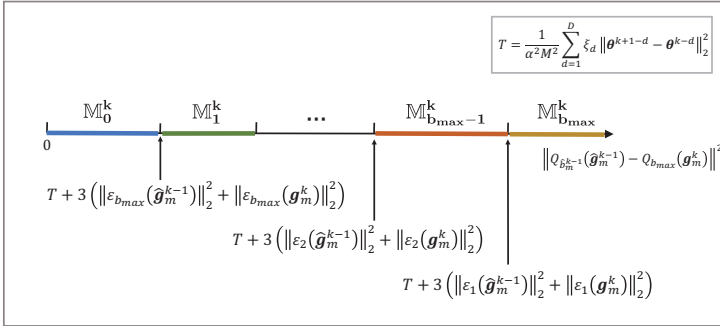


Fig. 3. The principle of the precision selection criterion.

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

For computation simplicity, a two-level variant of AQG is also proposed in this paper. 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 .

- 1: **for** $k = 1, 2, \dots, K$ **do**
- 2: Server broadcasts θ^k to all workers.
- 3: **for** each client $m \in \mathbb{M}$ **in parallel do**
- 4: Worker m computes \mathbf{g}_m^k and $Q_{b_{max}}(\mathbf{g}_m^k)$.
- 5: **if** (5) with $b = 1$ holds for worker m **then**
- 6: **for** $b = b_{max}, b_{max} - 1, \dots, 1$ **do**
- 7: **if** (5) with b holds for worker m **then**
- 8: Worker m computes and sends $Q_b(\mathbf{g}_m^k)$.
- 9: Set $b_m^k = b$.
- 10: Set $\hat{\mathbf{g}}_m^k = \mathbf{g}_m^k$ and $\hat{b}_m^k = b$ on both sides.
- 11: **Break.**
- 12: **end if**
- 13: **end for**
- 14: **else**
- 15: Worker m sends nothing.
- 16: Set $b_m^k = 0$,
- 17: Set $\hat{\mathbf{g}}_m^k = \hat{\mathbf{g}}_m^{k-1}$ and $\hat{b}_m^k = \hat{b}_m^{k-1}$ on both sides.
- 18: **end if**
- 19: **end for**
- 20: Server updates θ^{k+1} by $\theta^k - \alpha \sum_{m=1}^M Q_{\hat{b}_m^k}(\hat{\mathbf{g}}_m^k)$.
- 21: **end for**

3.2 Quantization Scheme

For better comparison, we adapt the quantization scheme used in LAQ algorithm[25]. The scheme quantizes the difference between the new gradient \mathbf{g}_m^k and the last quantized upload $Q_{\hat{b}_m^{k-1}}(\hat{\mathbf{g}}_m^{k-1})$:

$$\Delta = \mathbf{g}_m^k - Q_{\hat{b}_m^{k-1}}(\hat{\mathbf{g}}_m^{k-1}) \quad (7)$$

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 Fig. 4. By projecting every real number in this range to the closest quantized value, \mathbf{g}_m^k can be represented by $Q_b(\mathbf{g}_m^k)$ with b bits for each element instead of 32/64 bits by default.

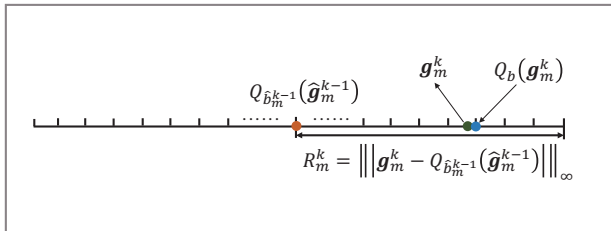


Fig. 4. Quantization scheme in AQG.

3.3 Augmented AQG for Client Dropouts

This paper 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 :

$$z_m^k \sim \text{Bernoulli}(p)$$

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

$$E[Q_{b_m^k}(\mathbf{g}_m^k)] = (1 - p) \cdot Q_{b_m^k}(\mathbf{g}_m^k) + p \cdot \mathbf{0} \quad (8)$$

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

In order to get the unbiased expectation, the upload is adjusted to $Q_{b_m^k}(\mathbf{g}_m^k)/(1 - p)$, and then:

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

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 $Q_{b_m^k}(\mathbf{g}_m^k)$ tends to be approximate to recent previous updates that may have been lost due to client dropouts.

Algorithm 2 Augmented AQG

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

Initialize: θ^1 .

```

1: for  $k = 1, 2, \dots, K$  do
2:   Server broadcasts  $\theta^k$  to all workers.
3:   for each client  $m \in \mathbb{M}$  in parallel do
4:     if  $z_m^k = 1$  then
5:       Worker  $m$  computes  $\mathbf{g}_m^k$  and  $Q_{b_{max}}(\mathbf{g}_m^k)$ .
6:       if (5) with  $b = 1$  holds for worker  $m$  then
7:         for  $b = b_{max}, b_{max} - 1, \dots, 1$  do
8:           if (5) with  $b$  holds for worker  $m$  then
9:             Worker  $m$  computes and sends  $Q_b(\mathbf{g}_m^k)$ .
10:            Set  $b_m^k = b$ .
11:            Set  $\hat{\mathbf{g}}_m^k = \mathbf{g}_m^k$  and  $\hat{b}_m^k = b$  on both sides.
12:            Break.
13:          end if
14:        end for
15:      end if
16:    else
17:      Worker  $m$  sends nothing.
18:      Set  $b_m^k = 0$ ,
19:      Set  $\hat{\mathbf{g}}_m^k = \hat{\mathbf{g}}_m^{k-1}$  and  $\hat{b}_m^k = \hat{b}_m^{k-1}$  on both sides.
20:    end if
21:  end for
22:  Server updates  $\theta^{k+1}$  by  $\theta^k - \alpha \sum_{m=1}^M Q_{\hat{b}_m^k}(\hat{\mathbf{g}}_m^k)$ .
23: end for

```

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

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

4 CONVERGENCE ANALYSIS

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

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

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

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

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

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:

$$\mathbb{V}(\theta^{k+1}) - \mathbb{V}(\theta^k) \leq 0 \quad (11)$$

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

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

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

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

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

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

4.2 Linear Convergence With Strongly-Convex Loss

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

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

Under Assumption 2, there is:

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

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

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

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 Assumption 1, Assumption 2 and Lemma 2, Lyapunov function and the quantization errors all converge at a linear rate:

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

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

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.

Table 2. Performance comparison of gradient-based algorithms.

Experiment setting		Iteration #	Communication #	Bit #	Transmission Reduction	
Logistic Regression	IID	Two Level AQG	500	3933	7952	41%
		Multilevel AQG	500	4372	8372	38%
		4-bit LAQ	500	3354	1.34×10^4	0
		4-bit QGD	500	9000	3.6×10^4	–*
	non-IID	Two Level AQG	500	4870	1.54×10^4	51%
		Multilevel AQG	500	8273	1.78×10^4	43%
		4-bit LAQ	500	7842	3.14×10^4	0
		32-bit GD ¹	500	9000	2.88×10^5	–
Neural Network	IID	Two Level AQG	2713	854	1708	34%
		Multilevel AQG	2881	974	1928	25%
		4-bit LAQ	2784	643	2572	0
		4-bit QGD	2890	28900	1.16×10^5	–
	non-IID	Two Level AQG	1319	1030	2060	44%
		Multilevel AQG	1702	977	1845	49%
		4-bit LAQ	2219	921	3684	0
		4-bit QGD	1251	12510	50040	–

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 against the baseline 4-bit LAQ.

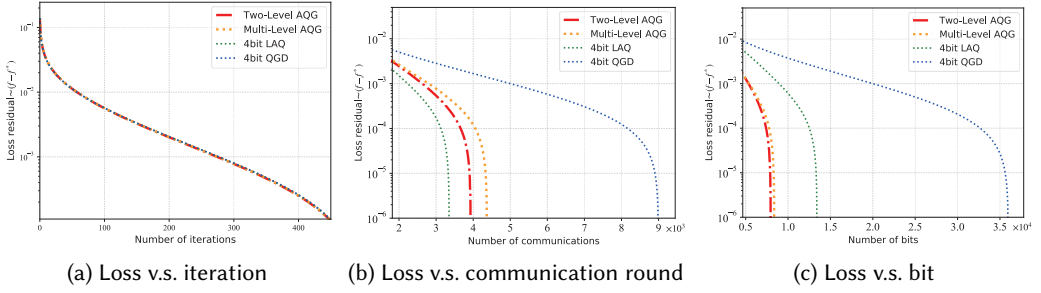


Fig. 5. Convergence of loss function with logistic regression and IID data distribution

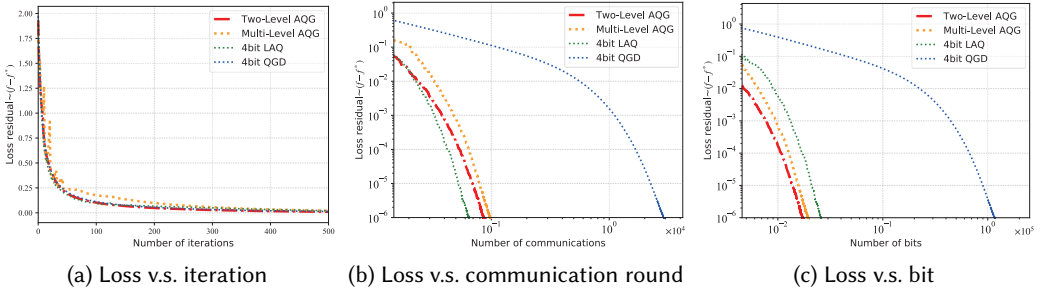


Fig. 6. Convergence of loss function with neural network and IID data distribution

5 EXPERIMENT RESULTS

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

5.1 Experimental Settings

For experiment simplicity, logistic regression is implemented with binary classification, and a fully connected network is built for non-convex optimization. The input and output dimension of the fully connected network is 784 and 10, respectively. For both Multi-level AQG and Tow-level AQG, the quantization bit number’s upper bound b_{max} is 4, the constant parameter D is 10, and the weights $\{\xi_d\}_{d=1}^D = 1/D$. Step size α is 0.008 for logistic regression and 0.02 for neural network.

In terms of datasets, both non-IID data distribution and IID data distribution are considered as follows:

non-IID Data Distribution: To simulate non-IID data distribution, a heterogeneous simulation dataset including 18 distributed data slices is used for logistic regression, and MNIST Dataset is used for multi-classification with the fully connected network by assigning each client with only one class of samples. The detailed description of the adopted dataset is provided in appendix. Obviously, the total client number M is set as 18 for logistic regression and 10 for fully connected network with these two datasets.

IID Data Distribution: For better comparison, the same binary classification dataset used to simulate non-IID data distribution is applied to simulate IID data distribution by uniformly distributing the samples across 18 clients. For the task with fully connected network, the MNIST dataset is distributed uniformly across 10 clients. Other parameters keep the same as in non-IID data distribution.

The experiment results are shown in Table. 2. For logistic regression, all algorithms run 500 iterations. For neural network, all algorithms run 4000 iterations, and we calculate the number of iteration, communication round and transmission bit 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. Thus, the higher the dimension of gradient is, the more significant transmission reduction AQG brings.

5.2 Performance of AQG with IID Data Distribution

With IID data distribution, training samples are distributed uniformly among clients. Fig. 5a shows that Multi-level AQG and the two-level variant of AQG both reach linear convergence rate as LAQ and QGD in strongly convex condition. Meanwhile, AQG significantly saves transmission bits compared against 4-bit LAQ and 4-bit QGD, as shown in Fig. 5c. It can be observed from Fig. 5b that the reduction of transmission bits is at the cost of a slight increase in communication rounds compared with LAQ, but it is worthy due to the significant reduction in overall transmission load.

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

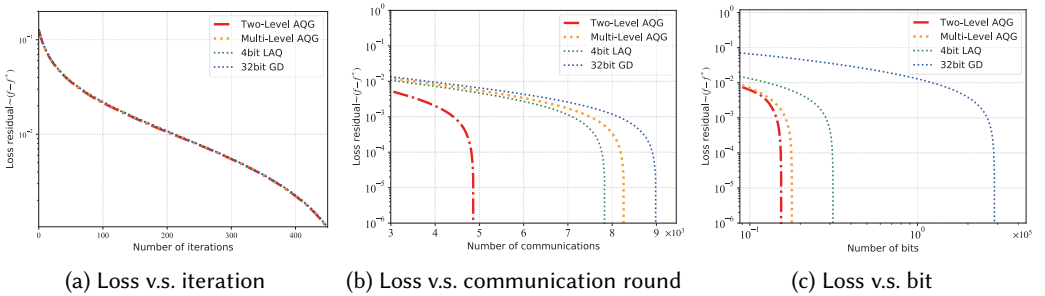


Fig. 7. Convergence of loss function with logistic regression and non-IID data distribution

5.3 Performance of AQG with non-IID Data Distribution

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

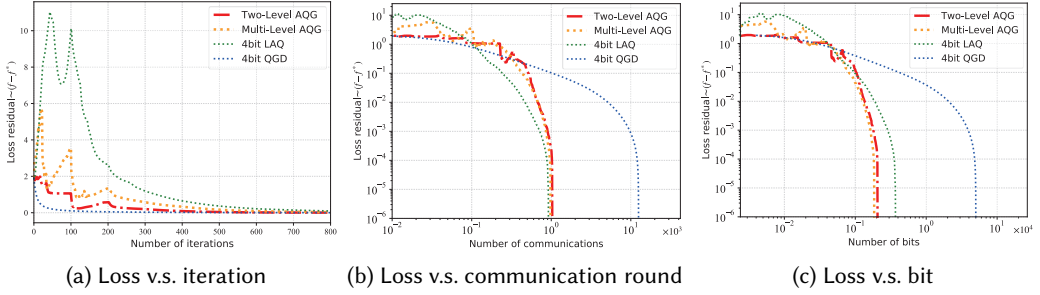


Fig. 8. Convergence of loss function with neural network and non-IID data distribution

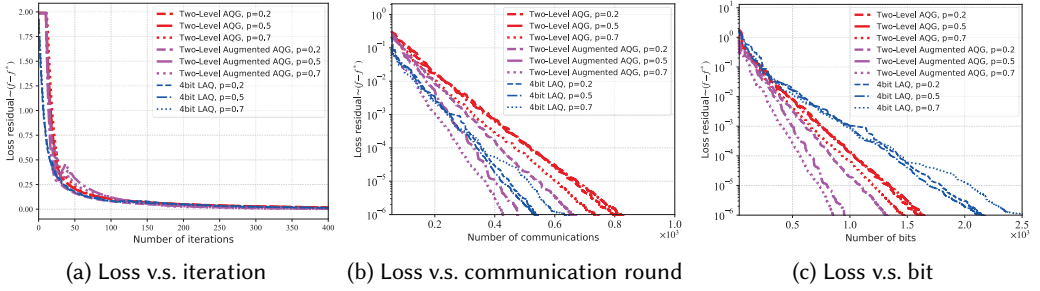


Fig. 9. Convergence of loss function with neural network ($p=0.2, 0.5$ and 0.7).

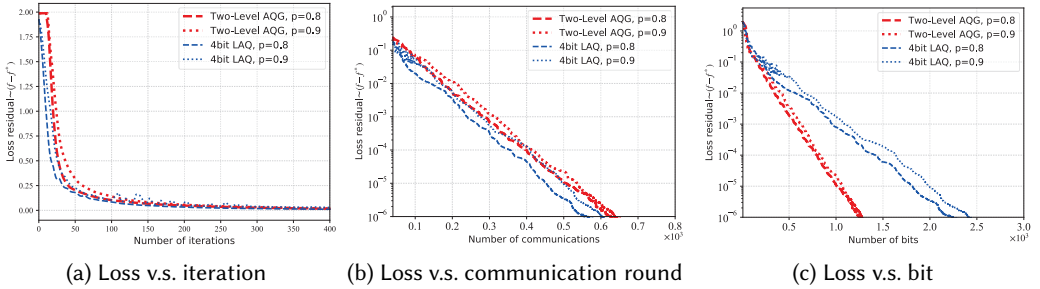


Fig. 10. Convergence of loss function with neural network ($p=0.8$ and 0.9).

5.4 Performance of AQG with client dropouts

In this part, we particularly focus on the setting of wireless network with mobile devices, where 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. Fig. 9 shows the performance of AQG with client dropping rate p as 0.2, 0.5 and 0.7. Experiment results demonstrate that both AQG and Augmented AQG require fewer transmission bits compared against LAQ. Meanwhile, Augmented AQG has a stronger ability to reduce transmission bits with the presence of such moderate client dropouts.

Fig. 10 shows the performance of AQQ with client dropping rate p as 0.8 and 0.9. Experiments show that AQQ manages to achieve stable convergence with ideal rates, and at the same time significantly reduces transmission bits even when there are only about 10% clients participating in gradient computation at each iteration. However, we notice that the augmented version of AQQ 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 Augmented AQQ no longer remains accurate and even induces more noise into the training. Thus, the Augmented AQQ is recommended to be applied in FL systems where the client dropping scale is moderate. Given the fact that the clients dropping rate is not likely to be so high in most practical systems, the augmented adaptive quantized gradient-based method is sufficient to address the dropping problem faced by FL.

6 CONCLUSION

This paper focuses on communication efficiency and the client dropout issue in FL, and proposes AQQ which not only adaptively adjusts the quantization level depending on 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 amount of bits, and vice versa. Since AQQ fully utilize the heterogeneousness of local data distribution to reduce unnecessary transmission, it achieves a larger transmission reduction with non-IID data distribution as expected. Compared against existing popular methods, AQQ leads to 25%-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%. Meanwhile, the Augmented AQQ brings extra transmission reduction with moderate-scale client dropouts commonly seen in practical scenarios, which indicates gradient amplification's effectiveness in suppressing the noise introduced by client dropouts.

Due to the aforementioned superiorities, AQQ can be used jointly with some other communication efficient methods for FL architectures, such as gradient sparsification[23], client selection based on local resources[13, 20, 29] and adaptively distributing subnetworks for heterogeneous clients [6, 9]. Such superiorities and flexibility endow great potentials for the proposed FL framework with AQQ. Future works include deploying AQQ jointly with such techniques in practical FL systems.

ACKNOWLEDGMENTS

This work is supported in part by Tsinghua-Foshan Innovation Special Fund (TFISF) under Grant No.2020THFS0109 and Guangdong Basic and Applied Basic Research Foundation under Grant No.2020A1515110887.

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). Association for Computing Machinery, New York, NY, USA, 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). 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 *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc.
- [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, Texas, USA). Association for Computing Machinery, New York, NY, USA, 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* (2020).
 - [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* (2018).
 - [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* (2019).
 - [9] Enmao Diao, Jie Ding, and Vahid Tarokh. 2020. HeteroFL: Computation and communication efficient federated learning for heterogeneous clients. *arXiv preprint arXiv:2010.01264* (2020).
 - [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] HA Güvenir, G. Demir?Z, and N. ?lter. 1998. Learning differential diagnosis of erythematous 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. 2018. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604* (2018).
 - [13] Yinghui He, Jinke Ren, Guanding Yu, and Jiantao Yuan. 2020. Resource Allocation for Wireless Federated Edge Learning based on Data Importance. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 1–6.
 - [14] Georgios A Kassis, 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.
 - [15] R. Kohavi. 1997. Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. 96 (1997), 202–207.
 - [16] Yujun Lin, Song Han, Huiji 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, Canada).
 - [17] 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. 13172–13179.
 - [18] 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.
 - [19] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agueria 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, USA, 1273–1282.
 - [20] Takayuki Nishio and Ryo Yonetani. 2019. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 1–7.
 - [21] 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 *INTERSPEECH*. 1058–1062.
 - [22] V. G. Sigillito, S. P. Wing, L. V. Hutton, and K. B. Baker. 1989. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest* 10, 3 (1989), 262–266.
 - [23] Navjot Singh, Deepesh Data, Jemin George, and Suhas Diggavi. 2020. SPARQ-SGD: Event-Triggered and Compressed Communication in Decentralized Optimization. In *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 3449–3456.
 - [24] 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* (Montréal, Canada). Curran Associates Inc., Red Hook, NY, USA, 4452–4463.
 - [25] J. Sun, T. Chen, G. B. Giannakis, Q. Yang, and Z. Yang. 2020. Lazily Aggregated Quantized Gradient Innovation for Communication-Efficient Federated Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1. <https://doi.org/10.1109/TPAMI.2020.3033286>
 - [26] 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, Canada) (*NIPS'18*). Curran Associates Inc., Red Hook, NY, USA, 9872–9883.
 - [27] Jianqiao Wang, 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, Canada). Curran Associates Inc., Red Hook, NY, USA, 1306–1316.

- [28] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–19.
- [29] Tongxin Zhu, Jianzhong Li, Zhipeng Cai, Yingshu Li, and Hong Gao. 2020. Computation scheduling for wireless powered mobile edge computing networks. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 596–605.

A MATHEMATICAL PROOF

A.1 Proof of Lemma 1

In AQG:

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

From the update rule of AQG:

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

From the definition of quantization error:

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

With inequality $\langle \mathbf{a}, \mathbf{b} \rangle \leq \frac{1}{2}\rho \|\mathbf{a}\|_2^2 + \frac{1}{2\rho} \|\mathbf{b}\|_2^2$ and (18):

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

Under Assumption 1:

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

$$+ \frac{\alpha}{2} \left\| \sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} [Q_{\hat{b}_m^k}(\hat{g}_m^k) - Q_{b_{\max}}(\hat{g}_m^k)] + \sum_{m \in \mathbb{M}_0^k} [Q_{\hat{b}_m^{k-1}}(\hat{g}_m^{k-1}) - Q_{b_{\max}}(\hat{g}_m^k)] \right\|_2^2 \quad (20)$$

The Lyapunov function of AQG is defined as:

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

Let $\beta_d = \frac{1}{\alpha} \sum_{j=d}^D \xi_j, \forall d \in \{1, \dots, D\}$, then:

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

Thus,

$$\begin{aligned} \mathbb{V}(\theta^{k+1}) - \mathbb{V}(\theta^k) &= f(\theta^{k+1}) - f(\theta^k) + \sum_{d=1}^D \beta_d \left\| \theta^{k+1-(d-1)} - \theta^{k-(d-1)} \right\|_2^2 - \sum_{d=1}^D \beta_d \left\| \theta^{k+1-d} - \theta^{k-d} \right\|_2^2 \\ &= f(\theta^{k+1}) - f(\theta^k) + \beta_1 \left\| \theta^{k+1} - \theta^k \right\|_2^2 + \sum_{d=1}^{D-1} (\beta_{d+1} - \beta_d) \left\| \theta^{k+1-d} - \theta^{k-d} \right\|_2^2 - \beta_D \left\| \theta^{k+1-D} - \theta^{k-D} \right\|_2^2 \\ &\leq -\alpha \left\langle \nabla f(\theta^k), \sum_{m=1}^M Q_{b_{\max}}(\hat{g}_m^k) \right\rangle + \frac{\alpha}{2} \left\| \nabla f(\theta^k) \right\|_2^2 \\ &\quad + \frac{\alpha}{2} \left\| \sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} [Q_{\hat{b}_m^k}(\hat{g}_m^k) - Q_{b_{\max}}(\hat{g}_m^k)] + \sum_{m \in \mathbb{M}_0^k} [Q_{\hat{b}_m^{k-1}}(\hat{g}_m^{k-1}) - Q_{b_{\max}}(\hat{g}_m^k)] \right\|_2^2 \\ &\quad + \left(\frac{L}{2} + \beta_1 \right) \left\| \theta^{k+1} - \theta^k \right\|_2^2 + \sum_{d=1}^{D-1} (\beta_{d+1} - \beta_d) \left\| \theta^{k+1-d} - \theta^{k-d} \right\|_2^2 - \beta_D \left\| \theta^{k+1-D} - \theta^{k-D} \right\|_2^2 \quad (23) \\ &= -\alpha \left\langle \nabla f(\theta^k), \sum_{m=1}^M Q_{b_{\max}}(\hat{g}_m^k) \right\rangle + \frac{\alpha}{2} \left\| \nabla f(\theta^k) \right\|_2^2 + \sum_{d=1}^{D-1} (\beta_{d+1} - \beta_d) \left\| \theta^{k+1-d} - \theta^{k-d} \right\|_2^2 - \beta_D \left\| \theta^{k+1-D} - \theta^{k-D} \right\|_2^2 \\ &\quad + \left(\frac{L}{2} + \beta_1 \right) \left\| \alpha \left\{ \sum_{m=1}^M Q_{b_{\max}}(\hat{g}_m^k) + \sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} [Q_{\hat{b}_m^k}(\hat{g}_m^k) - Q_{b_{\max}}(\hat{g}_m^k)] + \sum_{m \in \mathbb{M}_0^k} [Q_{\hat{b}_m^{k-1}}(\hat{g}_m^{k-1}) - Q_{b_{\max}}(\hat{g}_m^k)] \right\} \right\|_2^2 \\ &\quad + \frac{\alpha}{2} \left\| \sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} [Q_{\hat{b}_m^k}(\hat{g}_m^k) - Q_{b_{\max}}(\hat{g}_m^k)] + \sum_{m \in \mathbb{M}_0^k} [Q_{\hat{b}_m^{k-1}}(\hat{g}_m^{k-1}) - Q_{b_{\max}}(\hat{g}_m^k)] \right\|_2^2 \quad (24) \end{aligned}$$

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

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

$$+ \left(\frac{L}{2} + \beta_1\right)(1 + \rho_2)\alpha^2 \left\| \sum_{m=1}^M Q_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\|_2^2 \quad (25)$$

From $\left\| \sum_{i=1}^n \mathbf{a}_i \right\|_2^2 \leq n \sum_{i=1}^n \|\mathbf{a}_i\|_2^2$, there is:

$$\begin{aligned} & \left\| \sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} [Q_{\hat{b}_m^k}(\hat{\mathbf{g}}_m^k) - Q_{b_{\max}}(\hat{\mathbf{g}}_m^k)] + \sum_{m \in \mathbb{M}_0^k} [Q_{\hat{b}_m^{k-1}}(\hat{\mathbf{g}}_m^{k-1}) - Q_{b_{\max}}(\hat{\mathbf{g}}_m^k)] \right\|_2^2 \\ & \leq M \sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \left\| Q_{\hat{b}_m^k}(\hat{\mathbf{g}}_m^k) - Q_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\|_2^2 + M \sum_{m \in \mathbb{M}_0^k} \left\| Q_{\hat{b}_m^{k-1}}(\hat{\mathbf{g}}_m^{k-1}) - Q_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\|_2^2 \\ & = 2M \sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \left\| \varepsilon_{\hat{b}_m^k}(\hat{\mathbf{g}}_m^k) \right\|_2^2 + 2M \sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \left\| \varepsilon_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\|_2^2 + M \sum_{m \in \mathbb{M}_0^k} \left\| Q_{\hat{b}_m^{k-1}}(\hat{\mathbf{g}}_m^{k-1}) - Q_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\|_2^2 \end{aligned} \quad (26)$$

With (25) and (26):

$$\begin{aligned} \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}}(\hat{\mathbf{g}}_m^k) \right\rangle + \frac{\alpha}{2} \left\| \nabla f(\boldsymbol{\theta}^k) \right\|_2^2 \\ & + \left(\frac{L}{2} + \beta_1\right)(1 + \rho_2)\alpha^2 \left\| \sum_{m=1}^M Q_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\|_2^2 + \sum_{d=1}^{D-1} (\beta_{d+1} - \beta_d) \left\| \boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d} \right\|_2^2 - \beta_D \left\| \boldsymbol{\theta}^{k+1-D} - \boldsymbol{\theta}^{k-D} \right\|_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} [Q_{\hat{b}_m^k}(\hat{\mathbf{g}}_m^k) - Q_{b_{\max}}(\hat{\mathbf{g}}_m^k)] + \sum_{m \in \mathbb{M}_0^k} [Q_{\hat{b}_m^{k-1}}(\hat{\mathbf{g}}_m^{k-1}) - Q_{b_{\max}}(\hat{\mathbf{g}}_m^k)] \right\|_2^2 \\ & \leq -\alpha \left\langle \nabla f(\boldsymbol{\theta}^k), \sum_{m=1}^M Q_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\rangle + \frac{\alpha}{2} \left\| \nabla f(\boldsymbol{\theta}^k) \right\|_2^2 + \left(\frac{L}{2} + \beta_1\right)(1 + \rho_2)\alpha^2 \left\| \sum_{m=1}^M Q_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\|_2^2 \\ & + \sum_{d=1}^{D-1} (\beta_{d+1} - \beta_d) \left\| \boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d} \right\|_2^2 - \beta_D \left\| \boldsymbol{\theta}^{k+1-D} - \boldsymbol{\theta}^{k-D} \right\|_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}(\hat{\mathbf{g}}_m^k) \right\|_2^2 + \sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \left\| \varepsilon_{b_{\max}}(\hat{\mathbf{g}}_m^k) \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}}(\hat{\mathbf{g}}_m^{k-1}) - Q_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\|_2^2 \end{aligned} \quad (27)$$

With the precision selection criterion (5):

$$\begin{aligned} & M \sum_{m \in \mathbb{M}_0^k} \left\| Q_{\hat{b}_m^{k-1}}(\hat{\mathbf{g}}_m^{k-1}) - Q_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\|_2^2 \\ & \leq \frac{M^2}{\alpha^2 M^2} \sum_{d=1}^D \xi_d \left\| \boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d} \right\|_2^2 + 3M \sum_{m \in \mathbb{M}_0^k} \left(\left\| \varepsilon_{b_{\max}}(\hat{\mathbf{g}}_m^{k-1}) \right\|_2^2 + \left\| \varepsilon_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\|_2^2 \right) \end{aligned} \quad (28)$$

Thus,

$$\begin{aligned}
& \mathbb{V}(\boldsymbol{\theta}^{k+1}) - \mathbb{V}(\boldsymbol{\theta}^k) \\
& \leq -\alpha \left\langle \nabla f(\boldsymbol{\theta}^k), \sum_{m=1}^M Q_{b_{\max}}(\hat{\boldsymbol{g}}_m^k) \right\rangle + \frac{\alpha}{2} \left\| \nabla f(\boldsymbol{\theta}^k) \right\|_2^2 + \left(\frac{L}{2} + \beta_1 \right) (1 + \rho_2) \alpha^2 \left\| \sum_{m=1}^M Q_{b_{\max}}(\hat{\boldsymbol{g}}_m^k) \right\|_2^2 \\
& + \sum_{d=1}^{D-1} (\beta_{d+1} - \beta_d) \left\| \boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d} \right\|_2^2 - \beta_D \left\| \boldsymbol{\theta}^{k+1-D} - \boldsymbol{\theta}^{k-D} \right\|_2^2 \\
& + 2 \left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1 \right) (1 + \rho_2^{-1}) \alpha^2 \right] M \left(\sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \left\| \varepsilon_{\hat{b}_m^k}(\hat{\boldsymbol{g}}_m^k) \right\|_2^2 + \sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \left\| \varepsilon_{b_{\max}}(\hat{\boldsymbol{g}}_m^k) \right\|_2^2 \right) \\
& + \left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1 \right) (1 + \rho_2^{-1}) \alpha^2 \right] \frac{1}{\alpha^2} \sum_{d=1}^D \xi_d \left\| \boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d} \right\|_2^2 \\
& + 3 \left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1 \right) (1 + \rho_2^{-1}) \alpha^2 \right] M \sum_{m \in \mathbb{M}_0^k} \left(\left\| \varepsilon_{b_{\max}}(\hat{\boldsymbol{g}}_m^{k-1}) \right\|_2^2 + \left\| \varepsilon_{b_{\max}}(\hat{\boldsymbol{g}}_m^k) \right\|_2^2 \right) \tag{29}
\end{aligned}$$

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

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

$$\begin{aligned}
&\leq \left[-\frac{\alpha}{2} + \frac{\alpha\rho_1}{2} + (L + 2\beta_1)(1 + \rho_2)\alpha^2\right] \|\nabla f(\boldsymbol{\theta}^k)\|_2^2 + \left[\frac{\alpha}{2\rho_1} + (L + 2\beta_1)(1 + \rho_2)\alpha^2\right] \left\| \sum_{m=1}^M \varepsilon_{b_{\max}}(\hat{\boldsymbol{g}}_m^k) \right\|_2^2 \\
&+ \left\{ \left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1 + \rho_2^{-1})\alpha^2\right] \frac{1}{\alpha^2} \xi_D - \beta_D \right\} \|\boldsymbol{\theta}^{k+1-D} - \boldsymbol{\theta}^{k-D}\|_2^2 \\
&+ \sum_{d=1}^{D-1} \left\{ \left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1 + \rho_2^{-1})\alpha^2\right] \frac{1}{\alpha^2} \xi_d + \beta_{d+1} - \beta_d \right\} \|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2 \\
&+ 3\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1 + \rho_2^{-1})\alpha^2\right] M \sum_{m \in \mathbb{M}_0^k} \left(\|\varepsilon_{b_{\max}}(\hat{\boldsymbol{g}}_m^{k-1})\|_2^2 + \|\varepsilon_{b_{\max}}(\hat{\boldsymbol{g}}_m^k)\|_2^2 \right) \\
&+ 2\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1 + \rho_2^{-1})\alpha^2\right] M \left(\sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \|\varepsilon_{b_b^k}(\hat{\boldsymbol{g}}_m^k)\|_2^2 + \sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \|\varepsilon_{b_{\max}}(\hat{\boldsymbol{g}}_m^k)\|_2^2 \right) \quad (32)
\end{aligned}$$

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

$$-\frac{\alpha}{2} + \frac{1}{2}\alpha\rho_1 + (L + 2\beta_1)(1 + \rho_2)\alpha^2 \leq 0 \quad (33a)$$

$$\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1 + \rho_2^{-1})\alpha^2\right] \frac{\xi_D}{\alpha^2} - \beta_D \leq 0 \quad (33b)$$

$$\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1 + \rho_2^{-1})\alpha^2\right] \frac{\xi_d}{\alpha^2} + \beta_{d+1} - \beta_d \leq 0 \quad (33c)$$

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

$$\sum_{d=1}^D \xi_d \leq \min \left\{ \frac{1 - \rho_1}{4(1 + \rho_2)}, \frac{1}{2(1 + \rho_2^{-1})} \right\} \quad (34a)$$

$$\alpha \leq \min \left\{ \frac{2}{L} \left[\frac{1 - \rho_1}{4(1 + \rho_2)} - \sum_{d=1}^D \xi_d \right], \frac{2}{L} \left[\frac{1}{2(1 + \rho_2^{-1})} - \sum_{d=1}^D \xi_d \right] \right\} \quad (34b)$$

The above analysis indicates that there is no need to modify these two parameters involved in LAQ[25].

A.2 Proof of Lemma 2

Under Assumption 2:

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

$$\begin{aligned}
& + 3\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1 + \rho_2^{-1})\alpha^2\right]M \sum_{m \in \mathbb{M}_0^k} \left(\left\| \varepsilon_{b_{\max}}(\hat{\mathbf{g}}_m^{k-1}) \right\|_2^2 + \left\| \varepsilon_{b_{\max}}(\mathbf{g}_m^k) \right\|_2^2 \right) \\
& + 2\left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1 + \rho_2^{-1})\alpha^2\right]M \left(\sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \left\| \varepsilon_{\hat{b}_m^k}(\hat{\mathbf{g}}_m^k) \right\|_2^2 + \sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \left\| \varepsilon_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\|_2^2 \right) \quad (35)
\end{aligned}$$

Let c and B be defined as:

$$\begin{aligned}
c = \min_{d=1, \dots, D} \left\{ 2\mu \left[\frac{\alpha}{2} - \frac{\alpha\rho_1}{2} - (L + 2\beta_1)(1 + \rho_2)\alpha^2 \right], 1 - \left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1 + \rho_2^{-1})\alpha^2 \right] \frac{\xi_D}{\alpha^2\beta_D}, \right. \\
\left. 1 - \left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1 + \rho_2^{-1})\alpha^2 \right] \frac{\xi_d}{\alpha^2\beta_d} + \frac{\beta_{d+1}}{\beta_d} \right\} \quad (36a)
\end{aligned}$$

$$B = \max \left\{ \frac{\alpha}{2\rho_1} + (L + 2\beta_1)(1 + \rho_2)\alpha^2, 3M \left[\frac{\alpha}{2} + \left(\frac{L}{2} + \beta_1\right)(1 + \rho_2^{-1})\alpha^2 \right] \right\} \quad (36b)$$

Then:

$$\begin{aligned}
\mathbb{V}(\boldsymbol{\theta}^{k+1}) - \mathbb{V}(\boldsymbol{\theta}^k) & \leq -c \left[f(\boldsymbol{\theta}^k) - f(\boldsymbol{\theta}^*) + \sum_{d=1}^D \beta_d \left\| \boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d} \right\|_2^2 \right] \\
& + B \left\| \sum_{m=1}^M \varepsilon_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\|_2^2 + B \sum_{m \in \mathbb{M}_0^k} \left(\left\| \varepsilon_{b_{\max}}(\hat{\mathbf{g}}_m^{k-1}) \right\|_2^2 + \left\| \varepsilon_{b_{\max}}(\mathbf{g}_m^k) \right\|_2^2 \right) \\
& + B \left(\sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \left\| \varepsilon_{\hat{b}_m^k}(\hat{\mathbf{g}}_m^k) \right\|_2^2 + \sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \left\| \varepsilon_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\|_2^2 \right) \quad (37)
\end{aligned}$$

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

Thus,

$$\begin{aligned}
\mathbb{V}(\boldsymbol{\theta}^{k+1}) & \leq (1 - c)\mathbb{V}(\boldsymbol{\theta}^k) \\
& + B \left\| \sum_{m=1}^M \varepsilon_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\|_2^2 + B \sum_{m \in \mathbb{M}_0^k} \left(\left\| \varepsilon_{b_{\max}}(\hat{\mathbf{g}}_m^{k-1}) \right\|_2^2 + \left\| \varepsilon_{b_{\max}}(\mathbf{g}_m^k) \right\|_2^2 \right) \\
& + B \left(\sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \left\| \varepsilon_{\hat{b}_m^k}(\hat{\mathbf{g}}_m^k) \right\|_2^2 + \sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \left\| \varepsilon_{b_{\max}}(\hat{\mathbf{g}}_m^k) \right\|_2^2 \right) \quad (39)
\end{aligned}$$

A.3 Proof of Theorem 1

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

$$4BMP\tau_{b_{\max}}^2 + BMP \sum_{b=1}^{b_{\max}} \tau_{b_m^k}^2 \leq \sigma_2 - \sigma_1 \quad (40a)$$

$$\frac{24L^2}{\mu} + 18\tau_{b_{\max}-b_m^k}^2 + 3\tau_{b_{\max}}^2 \leq \sigma_2 \quad (40b)$$

$$\alpha \geq \frac{\mu}{4L^2M^2} \quad (40c)$$

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

$$\begin{aligned} \mathbb{V}(\boldsymbol{\theta}^{k+1}) &\leq \sigma_1 \mathbb{V}(\boldsymbol{\theta}^k) \\ &+ B \left\| \sum_{m=1}^M \varepsilon_{b_{\max}}(\hat{\boldsymbol{g}}_m^k) \right\|_2^2 + B \sum_{m \in \mathbb{M}_0^k} \left(\left\| \varepsilon_{b_{\max}}(\hat{\boldsymbol{g}}_m^{k-1}) \right\|_2^2 + \left\| \varepsilon_{b_{\max}}(\boldsymbol{g}_m^k) \right\|_2^2 \right) \\ &+ B \left(\sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \left\| \varepsilon_{b_m^k}(\hat{\boldsymbol{g}}_m^k) \right\|_2^2 + \sum_{b=1}^{b_{\max}} \sum_{m \in \mathbb{M}_b^k} \left\| \varepsilon_{b_{\max}}(\hat{\boldsymbol{g}}_m^k) \right\|_2^2 \right) \end{aligned} \quad (41)$$

$$\begin{aligned} &\leq \sigma_1 \sigma_2^{k-1} \mathbb{V}(\boldsymbol{\theta}^1) + 4BMP\tau_{b_{\max}}^2 \sigma_2^{k-1} \mathbb{V}(\boldsymbol{\theta}^1) + BMP \sum_{b=1}^{b_{\max}} \tau_{b_m^k}^2 \sigma_2^{k-1} \mathbb{V}(\boldsymbol{\theta}^1) \\ &= (\sigma_1 + 4BMP\tau_{b_{\max}}^2 + BMP \sum_{b=1}^{b_{\max}} \tau_{b_m^k}^2) \sigma_2^{k-1} \mathbb{V}(\boldsymbol{\theta}^1) \leq \sigma_2^k \mathbb{V}(\boldsymbol{\theta}^1) \end{aligned} \quad (42)$$

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

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

$$\begin{aligned} \left\| \nabla f_m(\boldsymbol{\theta}_1) - \nabla f_m(\boldsymbol{\theta}_2) \right\|_{\infty} &\leq \left\| \sum_{m=1}^M (\nabla f_m(\boldsymbol{\theta}_1) - \nabla f_m(\boldsymbol{\theta}_2)) \right\|_{\infty} \\ &= \left\| \nabla f(\boldsymbol{\theta}_1) - \nabla f(\boldsymbol{\theta}_2) \right\|_{\infty} \\ &\leq L \left\| \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\|_{\infty}, \quad \forall m \in \{1, \dots, M\} \end{aligned} \quad (43)$$

With (43) and the proposed precision selection criterion (5), there is:

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

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

$$\begin{aligned}
&\leq L\sqrt{2\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^*\|_2^2 + 2\|\boldsymbol{\theta}^* - \boldsymbol{\theta}^k\|_2^2} + \|\varepsilon_{b_{\max}}(\mathbf{g}_m^k)\|_\infty \\
&+ \sqrt{\frac{1}{\alpha^2 M^2} \sum_{d=1}^D \xi_d \|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2 + 3(\|\varepsilon_{b_{\max}-b_m^k}(\hat{\mathbf{g}}_m^{k-1})\|_\infty^2 + \|\varepsilon_{b_{\max}-b_m^k}(\mathbf{g}_m^k)\|_\infty^2)} \quad (46)
\end{aligned}$$

Under Assumption 2 with (13),

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

$$\begin{aligned}
&\leq \frac{12L^2}{\mu} \left[f(\boldsymbol{\theta}^{k+1}) - f(\boldsymbol{\theta}^*) + f(\boldsymbol{\theta}^k) - f(\boldsymbol{\theta}^*) + \frac{\mu}{4L^2\alpha^2 M^2} \sum_{d=1}^D \xi_d \|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2 \right] \\
&+ 18P\tau_{b_{\max}-b_m^k}^2 \sigma_2^{k-1} \mathbb{V}(\boldsymbol{\theta}^1) + 3P\tau_{b_{\max}}^2 \sigma_2^{k-1} \mathbb{V}(\boldsymbol{\theta}^1) \quad (48)
\end{aligned}$$

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

$$\begin{aligned}
&\|\nabla f_m(\boldsymbol{\theta}^{k+1}) - Q_{\hat{b}_m^{k-1}}(\hat{\mathbf{g}}_m^{k-1})\|_\infty^2 \\
&\leq \frac{12L^2}{\mu} \left[f(\boldsymbol{\theta}^{k+1}) - f(\boldsymbol{\theta}^*) + f(\boldsymbol{\theta}^k) - f(\boldsymbol{\theta}^*) + \sum_{d=1}^D \sum_{j=d}^D \frac{\xi_j}{\alpha} \|\boldsymbol{\theta}^{k+1-d} - \boldsymbol{\theta}^{k-d}\|_2^2 \right] \\
&+ 18P\tau_{b_{\max}-b_m^k}^2 \sigma_2^{k-1} \mathbb{V}(\boldsymbol{\theta}^1) + 3P\tau_{b_{\max}}^2 \sigma_2^{k-1} \mathbb{V}(\boldsymbol{\theta}^1) \\
&\leq \frac{12L^2}{\mu} [\mathbb{V}(\boldsymbol{\theta}^{k+1}) + \mathbb{V}(\boldsymbol{\theta}^k)] + 18P\tau_{b_{\max}-b_m^k}^2 \sigma_2^{k-1} \mathbb{V}(\boldsymbol{\theta}^1) + 3P\tau_{b_{\max}}^2 \sigma_2^{k-1} \mathbb{V}(\boldsymbol{\theta}^1) \\
&\leq \frac{24L^2}{\mu} \sigma_2^{k-1} \mathbb{V}(\boldsymbol{\theta}^1) + 18P\tau_{b_{\max}-b_m^k}^2 \sigma_2^{k-1} \mathbb{V}(\boldsymbol{\theta}^1) + 3P\tau_{b_{\max}}^2 \sigma_2^{k-1} \mathbb{V}(\boldsymbol{\theta}^1) \\
&= \left(\frac{24L^2}{\mu P} + 18\tau_{b_{\max}-b_m^k}^2 + 3\tau_{b_{\max}}^2 \right) P \sigma_2^{k-1} \mathbb{V}(\boldsymbol{\theta}^1) \leq P \sigma_2^k \mathbb{V}(\boldsymbol{\theta}^1) \quad (49)
\end{aligned}$$

Thus,

$$\|\varepsilon_b(\mathbf{g}_m^k)\|_\infty^2 \leq \tau_b^2 \|\nabla f_m(\boldsymbol{\theta}^{k+1}) - Q_{\hat{b}_m^{k-1}}(\hat{\mathbf{g}}_m^{k-1})\|_\infty^2 \leq P\tau_b^2 \sigma_2^k \mathbb{V}(\boldsymbol{\theta}^1) \quad (50)$$

B SIMULATION DATASETS

Three binary classification datasets listed in Table. 3 are used together in order to simulate non-IID data distribution as Chen *et al.* do in the evaluation of LAQ [7]. Specifically, The number of features is preprocessed to be equal to the minimal number of features among the total three datasets, and each dataset is uniformly distributed across six clients.

Table 3. The heterogeneous simulation datasets used for logistic regression.

Dataset	# features	# samples	client index
Adult fat[15]	113	1605	1,2,3,4,5,6
Ionosphere[22]	34	351	7,8,9,10,11,12
Derm[11]	34	358	13,14,15,16,17,18