A-LAQ: Adaptive Lazily Aggregated Quantized Gradient

Afsaneh Mahmoudi[1], José Mairton Barros Da Silva Júnior[1,2], Hossein S. Ghadikolaei[3], and Carlo Fischione[1]
[1]Network and Systems Engineering, Electrical Engineering and Computer Science
[1]KTH Royal Institute of Technology, Stockholm, Sweden
[1]emails: {afmb, jmbdsj, carlofi}@kth.se
[2] Princeton University, NJ, USA
[3] Ericsson, Stockholm, Sweden
[3]{email: hossein.shokri.ghadikolaei@ericsson.com}

***Abstract*—Federated Learning (FL) plays a prominent role in solving machine learning problems with data distributed across clients. In FL, to reduce the communication overhead of data between clients and the server, each client communicates the local FL parameters instead of the local data. However, when a wireless network connects clients and the server, the communication resource limitations of the clients may prevent completing the training of the FL iterations. Therefore, communication-efficient variants of FL have been widely investigated. Lazily Aggregated Quantized Gradient (LAQ) is one of the promising communication-efficient approaches to lower resource usage in FL. However, LAQ assigns a fixed number of bits for all iterations, which may be communication-inefficient when the number of iterations is medium to high or convergence is approaching. This paper proposes Adaptive Lazily Aggregated Quantized Gradient (A-LAQ), which is a method that significantly extends LAQ by assigning an adaptive number of communication bits during the FL iterations. We train FL in an energy-constraint condition and investigate the convergence analysis for A-LAQ. The experimental results highlight that A-LAQ outperforms LAQ by up to a 50% reduction in spent communication energy and an 11% increase in test accuracy.**

***Index Terms*—Federated learning, adaptive transmission, LAQ, communication bits, edge learning.**

I. Introduction

Federated Learning (FL) is a framework in which the clients train a centralized model by communicating their computed local models while data remains at each client [1]. FL has been widely studied because it preserves local data privacy and reduces communication overhead by avoiding data transmission. FL clients contribute to FL training by computing and sharing a local FL vector. However, computation and communication of such local vectors in large-scale FL require extensive communication resources [2]. Furthermore, the resources needed for FL training may be available in wired networks but not on wireless devices due to communication and energy resource constraints. Thus, we must minimize communication resource expenditure and get the most accurate training possible.

Many papers have recently focused on communication, computation, latency, and energy-efficient FL [3]–[7]. Authors in [3] have tried to minimize the system's total spent communication energy under a latency constraint and could reduce up to 59.5 % energy expenditure compared to the conventional FL. Reference [4] studied the joint power and resource allocation for ultra-reliable low-latency communication in vehicular networks and proposed a distributed approach based on FL to estimate the tail distribution of the queue lengths. Finally, authors of [5]–[7] have proposed a causal setting to jointly minimize the FL loss function and the overall resource consumption for training. Their results highlighted that joint design of communication protocols and FL are crucial for resource-efficient and accurate FL training.

Besides resource optimization, communication-efficient methods like quantization [8], [9], compression [10], and sparsification [11] can significantly reduce the communication overhead at each communication iteration. Adaptive methods have been recently noticed for communication-efficient FL training [12]–[15]. Authors in [12] have proposed an adaptive quantization strategy named AdaQuantFL by which they can change the quantization level in the stochastic quantization method to improve communication efficiency. Reference [13] has considered an adaptive quantization and sparsification scheme for uplink transmission facilitated by non-orthogonal multiple access. Authors in [14] have proposed an online learning scheme for determining the communication and computation trade-off. This trade-off is controlled by the degree of gradient sparsity obtained by the estimated sign of the objective function's derivative. Authors of [15] have proposed an adaptive gradient compression approach that improves communication efficiency by adjusting the compression rate according to the actual characteristics of each client.

Lazily aggregated quantized gradients (LAQ) method [16] is a novel framework that achieves the same linear convergence as the gradient descent in strongly convex set-ups. In addition, LAQ saves communication resources by using fewer transmitted bits at each communication iteration. However, LAQ considers a constant number of bits at each global and local FL transmission, which may not be communication-efficient enough.

In this paper, we significantly extend LAQ by considering an adaptive number of bits during the FL training to further improve communication and resource efficiency. The critical

factors in our proposed method are the descent behavior and the *diminishing return* rule [17] in FL training for L-smooth and convex loss functions. Due to the diminishing return rule, the accuracy improvement of the final model reduces with every new local and global communication iteration. Thus, we propose an adaptive LAQ, which we called A-LAQ, in which the FL training starts with a higher number of communication bits and adapts the bits as the communication between the server and clients continues. As the number of communication iterations increases, we propose that the number of bits can either decrease or stay the same. In A-LAQ, we assign more communication bits to the first communication iterations to minimize the quantization error at the beginning steps of training. After some communication iterations, we reduce the number of communication bits while facing a minor reduction in the loss function during training. We also develop a convergence analysis of FL with A-LAQ. The numerical results show that energy-constraint FL with A-LAQ outperforms FL with LAQ by up to a 50% reduction in spent communication energy and an 11% increase in test accuracy.

We organize the rest of this paper as the following. Section II describes the general system model and problem formulation. In Section III, we explain the solution approaches and convergence analysis for A-LAQ. Section IV shows some numerical results of A-LAQ and its performance compared to LAQ, and we conclude the paper in Section V.

Notation: Normal font w, bold font small-case $\boldsymbol{w}$, bold-font capital letter $\boldsymbol{W}$, and calligraphic font $\mathcal{W}$ denote scalar, vector, matrix, and set, respectively. We define the index set $[N] = \{1, 2, \ldots, N\}$ for any integer N. We denote by $\|\cdot\|$ the l_2-norm, by $\lceil . \rceil$ the ceiling value, by $|\mathcal{A}|$ the cardinality of set $\mathcal{A}$, by $[\boldsymbol{w}]_i$ the entry i of vector $\boldsymbol{w}$, by $\boldsymbol{w}^{\mathrm{T}}$ the transpose of $\boldsymbol{w}$, and $\mathbb{1}_x$ is an indicator function taking 1 if and only if x is true and takes 0 otherwise.

II. System Model and Problem Formulation

In this section, we represent the system model and the problem formulation. Consider a star network of M worker nodes that cooperatively solve a distributed training problem involving a loss function $f(\boldsymbol{w})$. Consider D as the whole dataset distributed among each worker $j \in [M]$ with D_j data samples. Let tuple $(\boldsymbol{x}_{ij}, y_{ij})$ denote data sample i of $|D_j|$ samples of worker node j and $\boldsymbol{w} \in \mathbb{R}^d$ denote the model parameter at the master node. Considering $\sum_{j=1}^{M} |D_j| = |D|$, and $j, j' \in [M]$, $j \neq j'$, we assume $D_j \cap D_{j'} = \emptyset$, and defining $\rho^j := |D_j|/|D|$, we formulate the following training problem

$$\boldsymbol{w}^* \in \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) = \sum_{j=1}^{M} \rho^j f^j(\boldsymbol{w}), \tag{1}$$

where $f^j(\boldsymbol{w}) := \sum_{i=1}^{|D_j|} f(\boldsymbol{w}; \boldsymbol{x}_{ij}, y_{ij})/|D_j|$.

A. LAQ Summary

In this part, we briefly summarize LAQ and its important parameters [16]. Considering the communication bits b, we define the quantization granularity $\tau := 1/(2^b - 1)$, the quantized version of each local gradient at the global communication iteration k as $\boldsymbol{q}^j(\boldsymbol{w}_k) = \text{Quant}(\nabla f^j(\boldsymbol{w}_k); b), j \in [M]$. Each local gradient is element-wise quantized by projecting to the closest point in a uniformly discretized d-dimensional grid with radius of $R_k^j = \|\nabla f^j(\boldsymbol{w}_k) - \boldsymbol{q}^j(\boldsymbol{w}_{k-1})\|_\infty$. We assume that all the workers participate in the training, each local loss function $f^j(\boldsymbol{w}_k)$ is L_j-smooth, the aggregated loss function $f(\boldsymbol{w}_k)$ is L-smooth and μ-strongly convex. Defining $\varepsilon_k^j := \nabla f^j(\boldsymbol{w}_k) - \boldsymbol{q}^j(\boldsymbol{w}_k)$ as the local quantization error, the aggregated quantization error is obtained as $\varepsilon_k := \sum_{j=1}^{M} \varepsilon_k^j$ and the aggregated quantized gradient is $\boldsymbol{q}_k := \sum_{j=1}^{M} \boldsymbol{q}^j(\boldsymbol{w}_k)$. The global updates in LAQ is $\boldsymbol{w}_k = \boldsymbol{w}_{k-1} - \alpha \nabla_{k-1}$, where $\tilde{\nabla}_k = \tilde{\nabla}_{k-1} + \sum_{j=1}^{M} \delta \boldsymbol{q}_k^j$ and $\delta \boldsymbol{q}_k^j := \boldsymbol{q}^j(\boldsymbol{w}_k) - \boldsymbol{q}^j(\boldsymbol{w}_{k-1})$.

B. Adaptive LAQ

In this subsection, we propose A-LAQ, in which we let b_k be the adaptive number of communication bits, and we introduce $\tau_k := 1/(2^{b_k} - 1)$ at each communication iteration $1 \leq k \leq K$. The global update in FL with A-LAQ is similar to LAQ, but the number of communication bits b_k becomes adaptive. First, we propose the following optimization problem, which formalizes the general scope of this paper:

$$\begin{aligned}
\underset{\boldsymbol{w}, k_0, K}{\text{minimize}} \quad & f(\boldsymbol{w}) && \text{(2a)}\\
\text{subject to} \quad & \boldsymbol{w}_k = \boldsymbol{w}_{k-1} - \alpha \tilde{\nabla}_{k-1}, \quad k = 1, \ldots, K && \text{(2b)}\\
& \tilde{\nabla}_k = \tilde{\nabla}_{k-1} + \sum_{j=1}^{M} \delta \boldsymbol{q}_k^j, \quad k = 1, \ldots, K && \text{(2c)}\\
& \delta \boldsymbol{q}_k^j = \boldsymbol{q}^j(\boldsymbol{w}_k) - \boldsymbol{q}^j(\boldsymbol{w}_{k-1}), \quad k = 1, \ldots, K && \text{(2d)}\\
& b_k = b^{\max} \mathbb{1}_{k \leq k_0} && \text{(2e)}\\
& \quad + b_0 \mathbb{1}_{k = k_0 + 1} + \lceil \eta_{k-1} b_{k-1} \rceil \mathbb{1}_{k > k_0 + 1}\\
& b_k \geq 2, \quad k = 1, \ldots, K && \text{(2f)}\\
& \sum_{k=1}^{K} E_k \leq E, \quad k = 1, \ldots, K, && \text{(2g)}
\end{aligned}$$

where $\boldsymbol{w}_k$ is the global FL parameter at each communication iteration k, $\rho^j, j \in [M]$ is the local weight, α is the step size, E_k is the communication energy spent at each communication iteration k, E is the total communication energy budget, and $k_0 \leq K$ is the number of the first communication iterations by which we assign $b_k = b^{\max}$, where $b^{\max}$ and b_0 are the given number of bits. We propose to update $b_k = \lceil \eta_{k-1} b_{k-1} \rceil$ for $k = \max\{3, k_0\}, \ldots, K$, by introducing η_{k-1} as

$$\eta_{k-1} := \min \left\{ \frac{\|f(\boldsymbol{w}_{k-1}) - f(\boldsymbol{w}_{k-2})\|}{\|f(\boldsymbol{w}_{k-2}) - f(\boldsymbol{w}_{k-3})\|}, 1 \right\}, \tag{3}$$

where the rationale of such a choice is the diminishing return rule. Constraints (2b)-(2d) reveal global LAQ update, constraints (2e) and (2f) show the adaptive b_k, and constraint (2g) is the overall communication energy limitation.

Optimization problem (2) aims to solve an FL problem in a communication energy-limited set-up. Although LAQ is a promising communication-efficient method, we show

that under the same resource limitation, A-LAQ saves more communication resources than LAQ. The set-up for A-LAQ is to assign a high number of communication bits to the communication iterations $1, \ldots, k_0$. Afterward, the training continues with b_0 communication bits, while $b_0 < b$ (where recall that b is the number of bits used by LAQ), and follows a non-increasing sequence of bits as implied by (3).

Optimization problem (2) is not practical because it requires K and the future local gradients for $k = 1, \ldots, K$ at the beginning of the training. Since it is impossible to have the information of local parameters and K beforehand, we call such a problem *non-causal* [5]. Therefore, in the rest of this paper, we focus on developing causal and practical solution approaches which do not need the future information of local gradients and K.

III. Solution Approach

This section provides a solution approach for optimization problem (2). Since optimization problem (2) is non-causal, we first calculate k_0, then proceed to calculate K and $\boldsymbol{w}^*$ in a causal way. To obtain k_0, we propose to solve a new optimization problem demonstrating the effect of the diminishing return rule on energy expenditure. After computing k_0, we simplify the optimization problem (2) and solve it to find K and $\boldsymbol{w}^*$ causally until the energy budget constraint is fulfilled.

A. Preliminary Results

To calculate k_0, we propose an optimization problem considering the diminishing return rule and energy expenditure. The idea behind A-LAQ is to change the number of communication bits to cope with the diminishing return rule. In other words, A-LAQ tries to associate a different number of communication bits at each communication iteration k to save the extra communication energy the clients spend before FL converges. Therefore, we define the energy-per-progress ratio function $E_f(\boldsymbol{w}_k, k; \mathcal{I}_k^j)$, where $\mathcal{I}_k^j$ is set of network's clients parameters, as

$$E_f(\boldsymbol{w}_k, k; M, [p_k^j]_j, [t_k^j]_j) := \frac{\sum_{k'=1}^{k} \sum_{j=1}^{M} p_{k'}^j t_{k'}^j}{f(\boldsymbol{w}_0) - f(\boldsymbol{w}_k)}, \ k \geq 1, \tag{4}$$

where $p_{k'}^j$ and $t_{k'}^j$ are respectively the transmission power and latency of each client $j \in [M]$ at every communication iteration $k' = 1, \ldots, k$. We assume that the client powers are constant at each communication iteration k', as $p_{k'}^j = p^j, j \in [M]$. Defining client transmission rate r^j bits/sec, we compute the transmission latency for each client $j \in [M]$, as $t_{k'}^j = b_{k'} d / r^j$ sec, where d is the dimension of the local and global parameters. Consider r^j as

$$r^j = \mathrm{BW}^j \log_2 \left(1 + \frac{p^j H^j}{N_0 \mathrm{BW}^j}\right), \tag{5}$$

where N_0 is the power spectrum density of noise, H^j is the channel gain and BW^j is the bandwidth allocated to each client $j \in [M]$. Defining power vector $\boldsymbol{p} := [p^1, \ldots, p^M]$, bit vector $\boldsymbol{b} := [b_1, \ldots, b_K]$, and the rate vector $\boldsymbol{r} := [r^1, \ldots, r^M]$, we have

$$E_f(\boldsymbol{w}_k, k; \boldsymbol{b}, M, \boldsymbol{p}, \boldsymbol{r}) = \frac{\sum_{k'=1}^{k} E_{k'}}{f(\boldsymbol{w}_0) - f(\boldsymbol{w}_k)} = \frac{\sum_{k'=1}^{k} b_{k'} \sum_{j=1}^{M} \frac{p^j d}{\mathrm{BW}^j \log_2(1 + \frac{p^j H^j}{N_0})}}{f(\boldsymbol{w}_0) - f(\boldsymbol{w}_k)}, \ k = 1, \ldots, K. \tag{6}$$

Now, considering $\boldsymbol{b} = b^{\max}\mathbf{1}$, we aim to minimize $E_f(\boldsymbol{w}_k, k; \boldsymbol{b}, M, \boldsymbol{p}, \boldsymbol{r})$ as

$$\underset{k, \boldsymbol{w}, K}{\text{minimize}} \quad E_f(\boldsymbol{w}_k, k; b^{\max}\mathbf{1}, M, \boldsymbol{p}, \boldsymbol{r}) \tag{7a}$$

$$\text{subject to} \quad \boldsymbol{w}_k = \boldsymbol{w}_{k-1} - \alpha \tilde{\nabla}_{k-1}, \quad k = 1, \ldots, K \tag{7b}$$

$$\tilde{\nabla}_k = \tilde{\nabla}_{k-1} + \sum_{j=1}^{M} \delta \boldsymbol{q}_k^j, \quad k = 1, \ldots, K \tag{7c}$$

$$\delta \boldsymbol{q}_k^j = \boldsymbol{q}^j(\boldsymbol{w}_k) - \boldsymbol{q}^j(\boldsymbol{w}_{k-1}), \quad k = 1, \ldots, K \tag{7d}$$

$$f(\boldsymbol{w}_k) = \sum_{j=1}^{M} \rho^j f^j(\boldsymbol{w}_k), \quad k = 1, \ldots, K, \tag{7e}$$

$$\sum_{k=1}^{K} E_k \leq E. \tag{7f}$$

To solve optimization problem (7), we propose the following Lemma, which demonstrates the conditions for discrete convexity [18] of $E_f(\boldsymbol{w}_k, k; b^{\max}, M, \boldsymbol{p}, \boldsymbol{r})$.

Lemma 1. *Let $f(\boldsymbol{w})$ be μ-strongly convex and L-smooth. Assume $b^{\max} = 32$ bits which represents the quantization full accuracy. Then, $E_f(\boldsymbol{w}_k, k; b^{\max}, M, \boldsymbol{p}, \boldsymbol{r})$ is discrete convex w.r.t. k.*

Proof: See Appendix A-A □

Lemma 1 demonstrates that $E_f(\boldsymbol{w}_k, k; b^{\max}, M, \boldsymbol{p}, \boldsymbol{r})$ has a unique minimum w.r.t. k. Thus, we calculate k_0 as

$$k_0 \in \underset{k \in \mathbb{N}}{\arg\min} \ E_f(\boldsymbol{w}_k, k; b^{\max}, M, \boldsymbol{p}, \boldsymbol{r}) \tag{8a}$$

$$\text{subject to} \quad \text{(7b)} - \text{(7f)}. \tag{8b}$$

After computing k_0, we re-write the optimization problem (2) as

$$\underset{\boldsymbol{w}, K, \boldsymbol{b}}{\text{minimize}} \quad f(\boldsymbol{w}) \tag{9a}$$

$$\text{subject to} \quad b_k = b_0 \mathbb{1}_{k=k_0+1} + \lceil \eta_{k-1} b_{k-1} \rceil \mathbb{1}_{k>k_0+1} \tag{9b}$$

$$b_k \geq 2, \quad k = k_0 + 1, \ldots, K \tag{9c}$$

$$\sum_{k=k_0+1}^{K} E_k \leq E - \sum_{k=1}^{k_0} E_k \tag{9d}$$

$$\text{(2b)} - \text{(2d)}. \tag{9e}$$

Now, equipped with the preliminary results of this subsection, we are ready to solve optimization problem (2) in the following subsection.

B. Solution Approach

First, we consider Lemma 1 and compute k_0 according to the following proposition.

Proposition 1. *Let $f(\boldsymbol{w})$ be μ-strongly convex and L-smooth. Consider $b^{\max} = 32$ bits. Thus, $k_0 = \min\{k_e, k_f\}$, where*

$$k_e := \textit{the first value of } k \textit{ such that } E_k > E - \sum_{k'=1}^{k-1} E_{k'}, \quad (10)$$

and

$$k_f := \textit{the first value of } k \textit{ such that} \quad (11)$$
$$k < \frac{f(\boldsymbol{w}_0) - f(\boldsymbol{w}_k)}{f(\boldsymbol{w}_{k-1}) - f(\boldsymbol{w}_k)}.$$

Proof: See Appendix A-B □

Note that when $k_0 = k_e$, constraint (2g) is fulfilled, thus the training is complete and $K = k_0$, $\boldsymbol{b} = b^{\max}\mathbf{1}$. Otherwise, after computing k_0, we focus on optimization problem (9) to obtain K, $\boldsymbol{w}$ and $\boldsymbol{b}$. Considering the non-increasing sequence of b_k for $k \in [k_0+1, K]$ in (9b) and (9c) along with the energy constraint of (9d), we obtain

$$\left(32k_0 + b_0 + \sum_{k'=k_0+2}^{K} b_{k'}\right) \sum_{j=1}^{M} \frac{p^j d}{\mathrm{BW}^j \log_2(1 + \frac{p^j H^j}{N_0})} \leq E. \quad (12)$$

Eq. (12) plays a critical role in FL training for the communication iteration $k \geq k_0 + 1$. It means that K is obtained while the energy budget E is spent. The following lemma determines when we can terminate the FL with A-LAQ training by finding K.

Lemma 2. *Let $f(\boldsymbol{w})$ be μ-strongly convex and L-smooth and $b^{\max} = 32$ bits. For any $k > k_0$, we obtain $K = k$ if*

$$\eta_k b_k \sum_{j=1}^{M} \frac{p^j d}{BW^j \log_2(1 + \frac{p^j H^j}{N_0})} > E - \sum_{k'=1}^{k} E_k. \quad (13)$$

Proof: See Appendix A-C □

Therefore, the FL training with A-LAQ continues until K is obtained. Algorithm 1 summarizes all the steps for FL with A-LAQ.

Theorem 1. *Let $f(\boldsymbol{w})$ be μ-strongly convex and L-smooth. Assume $b^{\max} = 32$ and $b_0 < b$ be given. Then, by solving optimization problems* (7) *and* (9)*, we achieve an exact solution for optimization problem* (2)*.*

Proof: In this paper, we propose to solve optimization problem (2) in a causal way. Thus, we first have to compute k_0 to determine when we must adapt the number of bits. To do so, we propose to solve optimization problem (7) which highlights the diminishing return rule and energy expenditure. The solution to (7) is exact and mathematically calculated by either (10) or (11). Next, calculate K and $\boldsymbol{w}$, which is another causal approach, and the exact solution for K is obtained by (13). □

Algorithm 1: Federated Learning with A-LAQ

Inputs: $\boldsymbol{w}_0$, M, $(\boldsymbol{x}_{ij}, y_{ij})_{i,j}$, α, $b^{\max}$, b_0, $\boldsymbol{r}$, $\boldsymbol{p}$, $\{|D_j|\}_{j\in[M]}$, $\{\rho^j\}_{j\in[M]}$, μ, L.
Initialize: $\tilde{\nabla}_0$, $K = +\infty$, $k_0 = k_e = k_f = 0$, $(b_k)_{k\in[K]} = b^{\max}$
Master node broadcasts $\boldsymbol{w}_0$ to all nodes
while $K = +\infty$ **do**
  for $k = 1, \ldots, K$ **do**
    for $j \in [M]$ **do**
      Calculate $\nabla f^j(\boldsymbol{w}_k)$, $\boldsymbol{q}^j(\boldsymbol{w}_k)$, $\delta\boldsymbol{q}_k^j$ and $f^j(\boldsymbol{w}_k)$
      Send $\delta\boldsymbol{q}_k^j$ and $f^j(\boldsymbol{w}_k)$ to the master node
    end for
    Wait until master node collects all $\{\delta\boldsymbol{q}_k^j\}_{j\in[M]}$ and update $f(\boldsymbol{w}_k)$, and $\tilde{\nabla}_k$ and $\boldsymbol{w}_k$ according to (2b), (2c)
    if $k_0 = 0$ **then**
      if $\max\{k_f, k_e\} > 0$ **then**
        Set $k_0 = k$
        Set $b_{k+1} = b_0$
      end if
    else
      Calculate η_k according to (3)
      Set $b_{k+1} = \lceil \eta_k b_k \rceil$
      if Inequality (13) is true **then**
        Set K = k
      end if
    end if
    Set $k \leftarrow k + 1$
  end for
end while
Return $\boldsymbol{w}_K$, k_0, K, $(b_k)_{k\in[K]}$

C. Convergence Analysis

In this subsection, we investigate the convergence of A-LAQ. Since $\|\varepsilon_k^j\|_\infty \leq \tau_k R_k^j$, for each element of $[\varepsilon_k^j]_i, i = 1, \ldots, d$, we have $|[\varepsilon_k^j]_i| \leq \tau_k R_k^j$, thus

$$\|\varepsilon_k^j\|_2 \leq \sqrt{d}\,\tau_k R_k^j. \quad (14)$$

According to definition of ε_k in LAQ, $\varepsilon_k = \sum_{j=1}^M \varepsilon_k^j$, thus

$$\|\varepsilon_k\|_2 = \left\|\sum_{j=1}^{M} \varepsilon_k^j\right\|_2 \overset{\text{triangle}}{\leq} \sum_{j=1}^{M} \|\varepsilon_k^j\|_2 \overset{(14)}{\leq} \sum_{j=1}^{M} \sqrt{d}\,\tau_k R_k^j. \quad (15)$$

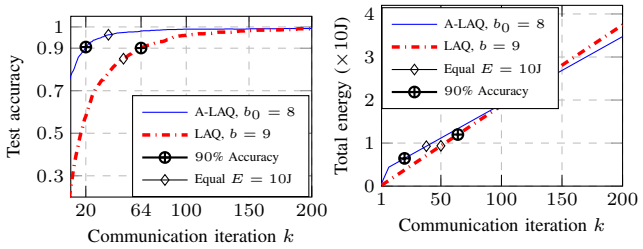
Then, considering the inequalities (14) and (15), and for every b_k, we give the following proposition.

Proposition 2. *Let $f(\boldsymbol{w})$ be μ-strongly convex and L-smooth, and $f^* := f(\boldsymbol{w}^*)$ be the loss function value of the optimal solution of optimization problem* (1)*. We define a Lyapunov function as*

$$\mathbb{V}(\boldsymbol{w}_k) := f(\boldsymbol{w}_k) - f^* + \sum_{i=1}^{k_1}\sum_{h=i}^{k_1} \frac{\zeta_h}{\alpha}\|\boldsymbol{w}_{k+1-i} - \boldsymbol{w}_{k-i}\|_2^2 + \gamma \sum_{j=1}^{M} \|\varepsilon_k^j\|_\infty^2, \quad (16)$$

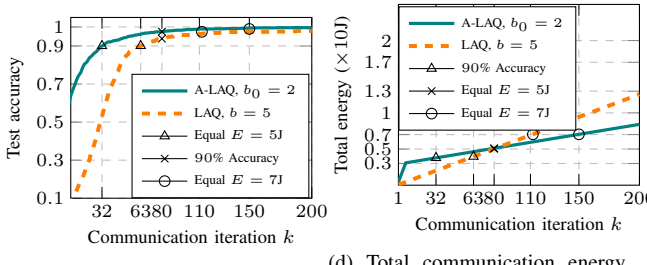
where $\zeta_h = \zeta, h \in [k_1]$ and γ are non-negative constants and $k_1 \leq k$. By $0 < \rho < 1$, $\beta_i - \beta_{i+1} = \beta_{k_1}, i = 1, \ldots, k_1 - 1$, $a \in (0, 1]$, $\alpha = a/L$, and $\gamma \geq d\alpha^2\left(L + 2\beta_1 + (2\rho\alpha)^{-1}\right)$, $\zeta < M/6\tau_{k+1}^2 dk_1$, and

$$\beta_{k_1} \geq \frac{dL + \frac{d}{2\alpha\rho}}{\frac{M}{3\tau_{k+1}^2\zeta} - 2dk_1}.$$



(a) Test accuracy, $M = 50$.

(b) Total communication energy, $M = 50$.



(c) Test accuracy, $M = 30$.

(d) Total communication energy, $M = 30$.

Fig. 1: Comparison of A-LAQ and LAQ for a) Test accuracy, and b) Total communication energy for $M = 50$, $b = 9$, $b_0 = 8$, $b^{\max} = 32$ and $k_0 = 7$. c) Test accuracy, and b) Total communication energy for $M = 30$, $b = 5$, $b_0 = 2$, $b^{\max} = 32$ with $k_0 = 7$.

Then, Lyapunov function (16) *is non-increasing, i.e.* $\mathbb{V}(\boldsymbol{w}_{k+1}) \leq \mathbb{V}(\boldsymbol{w}_k), k \geq 1$.

Proof: See Appendix A-D. □

Proposition 2 shows that by proper choice of the Lyapunov function parameters, FL with A-LAQ converges.

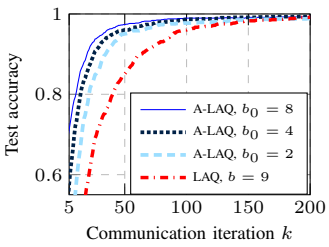
IV. Numerical Results

In this section, we illustrate our results from the previous sections and numerically show the extensive impact of A-LAQ on FL training. We consider solving a convex regression problem over a wireless network using a real-world dataset. To this end, we extract a binary dataset from MNIST (handwritten digits) by keeping only samples of digits 0 and 1 and then setting their labels to -1 and +1, respectively. We then randomly split the resulting dataset of 12600 samples among M worker nodes, each having $\{(\boldsymbol{x}_{ij}, y_{ij})\}$, where $\boldsymbol{x}_{ij} \in \mathbb{R}^{784}$ is a data sample i, which is a vectorized image at node $j \in [M]$ with corresponding digit label $y_{ij} \in \{-1, +1\}$. We use the following training loss function [19]

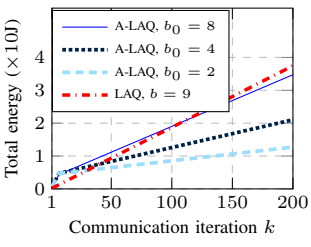
$$f(\boldsymbol{w}) = \sum_{j=1}^{M} \rho^j \sum_{i=1}^{|D_j|} \frac{1}{|D_j|} \log\left(1 + e^{-\boldsymbol{w}^T \boldsymbol{x}_{ij} y_{ij}}\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2, \tag{17}$$

where $\lambda \in (0, 1)$ is a given regularization parameter and each worker node $j \in [M]$ has the same number of samples, namely $|D_j| = |D_i| = |D|/M, \forall i, j \in [M]$.

We consider OFDMA for the uplink in a single cell system with the coverage radius of $\ell_c = 1$ Km. There are L_p cellular links on S_c subchannels. We model the subchannel power gain $h_l^s = \phi/(\ell^j)^3$, where ℓ^j is the distance between each client to the master node, following the Rayleigh fading, where ϕ has an exponential distribution with unitary mean. We consider



(a) Test accuracy.



(b) Total communication energy.

Fig. 2: Comparison between LAQ with $b = 9$, and A-LAQ with $b_0 = 2$, 4 and 8 for $M = 50$. a) Test accuracy shows that all three A-LAQ scenarios outperform LAQ. b) A-LAQ with smaller b_0 performs better in an energy limited FL.

the noise power in each subchannel as -170 dBm/Hz and the maximum transmit power of each link as 23 dBm. We assume that $S_c = 64$ subchannels, the total bandwidth of 10 MHz, and the subchannel bandwidth of 150 KHz.

Fig. 1 illustrates A-LAQ performance and compares it with LAQ. Figs. 1(a) and 1(b) show test accuracy for $M = 50$, $b = 9$, $b_0 = 8$, $b^{\max} = 32$ and $k_0 = 7$ is obtained. Each pair of black marks demonstrates the comparison between A-LAQ and LAQ either for the same energy budget E or the same test accuracy. For $E = 10$J, we obtain $K = 38$ for A-LAQ with test accuracy of 96% , and $K = 50$ for LAQ, with test accuracy of 85%. Besides, we observe that for achieving a test accuracy of 90%, A-LAQ spends 50% less energy and requires a smaller K than LAQ.

Figs. 1(c) and 1(d) address the test accuracy and total spent communication energy for $M = 30$, $b = 5$, $b_0 = 2$, $b^{\max} = 32$ with $k_0 = 7$. Similar to the previous arguments, for an equal test accuracy of 90%, A-LAQ outperforms LAQ by spending approximately the same energy but smaller K. For an energy budget $E = 5$J, A-LAQ and LAQ calculate the same K, but the test accuracy for A-LAQ is 4% higher than LAQ. We also observe that for $k \geq 80$, the total spent communication energy in A-LAQ is lower than LAQ, while the test accuracy of LAQ and A-LAQ are quite similar. Thus, when high communication energy resources are available, A-LAQ requires lower communication energy than LAQ to perform K iterations.

Fig. 2 compares A-LAQ performance of test accuracy and total communication energy for $M = 50$, with different values of $b_0 = 8$, 4, and 2. Fig. 2(a) shows test accuracy, and we observe that LAQ has the lowest value of test accuracy for all iterations. Fig. 2(b) demonstrates the total communication energy, which A-LAQ with $b_0 = 2$ and $b_0 = 5$, spends lower energy, while having very close test accuracy to A-LAQ with $b_0 = 8$. We conclude that A-LAQ with smaller b_0 outperforms A-LAQ with higher b_0 in terms of energy expenditure and test accuracy for the same energy budget.

V. Conclusion

In this paper, we considered Federated Learning and the LAQ algorithm and proposed an adaptive transmission framework, A-LAQ, by significantly extending LAQ. Different from

LAQ, A-LAQ used an adaptive number of communication bits in a communication energy-limited situation. We analyzed the convergence of A-LAQ, and we showed that A-LAQ could achieve a better performance in test accuracy (by an 11% increase) while reducing the communication energy by 50%.

Future Work: Our future work involves extending A-LAQ to communication-efficient scenarios with the best client selection policy. Also, we will consider the computation energy of clients and obtain the optimal sequences of bits to achieve a communication-computation energy-efficient A-LAQ.

REFERENCES

[1] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[2] H. Hellström, J. M. B. da Silva Jr, M. M. Amiri, M. Chen, V. Fodor, H. V. Poor, C. Fischione *et al.*, "Wireless for Machine Learning: A Survey," *Foundations and Trends® in Signal Processing*, vol. 15, no. 4, pp. 290–399, 2022.

[3] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient Federated Learning over wireless communication networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1935–1949, 2021.

[4] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Distributed Federated Learning for ultra-reliable low-latency vehicular communications," *IEEE Transactions on Communications*, vol. 68, no. 2, pp. 1146–1159, 2020.

[5] A. Mahmoudi, H. S. Ghadikolaei, and C. Fischione, "Cost-efficient distributed optimization in machine learning over wireless networks," in *IEEE International Conference on Communications (ICC)*, 2020.

[6] ——, "Machine learning over networks: Co-design of distributed optimization and communications," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2020.

[7] A. Mahmoudi, H. S. Ghadikolaei, J. M. B. Da Silva, and C. Fischione, "FedCau: A proactive stop policy for communication and computation efficient Federated Learning," *arXiv preprint arXiv:2204.07773*, 2022.

[8] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Federated Learning with quantized global model updates," *arXiv preprint arXiv:2006.10672*, 2020.

[9] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "UVeQFed: Universal vector quantization for Federated Learning," *IEEE Transactions on Signal Processing*, vol. 69, pp. 500–514, 2021.

[10] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient Federated Learning from non-i.i.d. data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3400–3413, 2020.

[11] S. Li, Q. Qi, J. Wang, H. Sun, Y. Li, and F. R. Yu, "GGS: General gradient sparsification for Federated Learning in edge computing," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–7.

[12] D. Jhunjhunwala, A. Gadhikar, G. Joshi, and Y. C. Eldar, "Adaptive quantization of model updates for communication-efficient Federated Learning," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3110–3114.

[13] H. Sun, X. Ma, and R. Q. Hu, "Adaptive Federated Learning with gradient compression in uplink NOMA," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 16 325–16 329, 2020.

[14] P. Han, S. Wang, and K. K. Leung, "Adaptive gradient sparsification for efficient Federated Learning: An online learning approach," in *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*, 2020, pp. 300–310.

[15] W. Yang, Y. Yang, X. Dang, H. Jiang, Y. Zhang, and W. Xiang, "A novel adaptive gradient compression approach for communication-efficient Federated Learning," in *2021 China Automation Congress (CAC)*, 2021, pp. 674–678.

[16] J. Sun, T. Chen *et al.*, "Lazily Aggregated Quantized Gradient (LAQ) innovation for communication-efficient federated learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2031–2044, 2022.

[17] N. C. Thompson, K. Greenewald *et al.*, "Deep learning's diminishing returns: The cost of improvement is becoming unsustainable," *IEEE Spectrum*, vol. 58, no. 10, pp. 50–55, 2021.

[18] B. L. Miller, "On minimizing nonseparable functions defined on the integers with an inventory application," *SIAM Journal on Applied Mathematics*, vol. 21, no. 1, pp. 166–185, 1971.

[19] K. Koh, S.-J. Kim, and S. Boyd, "An interior-point method for large-scale ℓ_1-regularized logistic regression," *Journal of Machine Learning Research*, vol. 8, no. Jul, pp. 1519–1555, 2007.

APPENDIX A

A. Proof of Lemma 1

This proof is ad-absurdum. Assume that The sequences of $E_f(\boldsymbol{w}_k, k; \mathcal{I}_k^j)$ is not discrete convex. Therefore, there is a $k > 1$ such that $E_f(\boldsymbol{w}_k, k; \mathcal{I}_k^j) > E_f(\boldsymbol{w}_{k-1}, k-1; \mathcal{I}_{k-1}^j)$ and $E_f(\boldsymbol{w}_k, k; \mathcal{I}_k^j) > E_f(\boldsymbol{w}_{k+1}, k+1; \mathcal{I}_{k+1}^j)$. According to the statement of Lemma 1, since $b_k = b^{\max}$, $k \leq k_0$, we consider $\sum_{k'=1}^{k} E_{k'} = kE_1$. Besides, $f(\boldsymbol{w})$ is μ-strongly convex and L-smooth, which means the sequence of $f(\boldsymbol{w}_k)$ have the descent behavior w.r.t. k, and satisfies $f(\boldsymbol{w}_k) - f(\boldsymbol{w}_{k+1}) \leq f(\boldsymbol{w}_{k-1} - f(\boldsymbol{w}_k)$. According to the definition of $E_f(\boldsymbol{w}_k, k; \mathcal{I}_k^j) = kE_1/(f(\boldsymbol{w}_0) - f(\boldsymbol{w}_k))$, we have $f(\boldsymbol{w}_0) - f(\boldsymbol{w}_k) = f(\boldsymbol{w}_0) - f(\boldsymbol{w}_{k-1}) + f(\boldsymbol{w}_{k-1}) - f(\boldsymbol{w}_k) \geq f(\boldsymbol{w}_0) - f(\boldsymbol{w}_{k-1})$, which means that both numerator and denominator of $E_f(\boldsymbol{w}_{k+1}, k+1; \mathcal{I}_{k+1}^j)$ are non-decreasing w.r.t. k. Now, if we assume that $E_f(\boldsymbol{w}_k, k; \mathcal{I}_k^j) > E_f(\boldsymbol{w}_{k-1}, k-1; \mathcal{I}_{k-1}^j)$ and $E_f(\boldsymbol{w}_k, k; \mathcal{I}_k^j) > E_f(\boldsymbol{w}_{k+1}, k+1; \mathcal{I}_{k+1}^j)$, it results in a decrease in the denominator from k to $k+1$, thus we obtain that $f(\boldsymbol{w}_0) - f(\boldsymbol{w}_k) \geq f(\boldsymbol{w}_0) - f(\boldsymbol{w}_{k+1})$ which is in contradiction with the behavior of $f(\boldsymbol{w}_k)$. Therefore, we conclude that $E_f(\boldsymbol{w}_k, k; \mathcal{I}_k^j)$ is discretely convex.

B. Proof of Proposition 1

First consider that $k_0 = k_e$, it means that the energy budget is determining k_0. As we mentioned in A-A, $E_k = E_0$ and $\sum_{k'=1}^{k} E_{k'} = kE_0$. Thus, when $E_0 > E - kE_0$, it results in energy limitation and then $k_0 = K = k$.

Next, consider that $k_0 = k_f$, according to Lemma 1, $E_f(\boldsymbol{w}_k, k; \mathcal{I}_k^j)$ is discrete convex and we obtain $k_0 = k$ when $E_f(\boldsymbol{w}_k, k; \mathcal{I}_k^j) - E_f(\boldsymbol{w}_{k-1}, k-1; \mathcal{I}_{k-1}^j) > 0$, see [5]. Thus,

$$
\begin{aligned}
&E_f(\boldsymbol{w}_k, k; \mathcal{I}_k^j) - E_f(\boldsymbol{w}_{k-1}, k-1; \mathcal{I}_{k-1}^j) = \\
&\frac{kE_0}{f(\boldsymbol{w}_0) - f(\boldsymbol{w}_k)} - \frac{(k-1)E_0}{f(\boldsymbol{w}_0) - f(\boldsymbol{w}_{k-1})} = \\
&\frac{kE_0}{f(\boldsymbol{w}_0) - f(\boldsymbol{w}_k)} - \frac{(k-1)E_0}{f(\boldsymbol{w}_0) - f(\boldsymbol{w}_{k-1})} > 0, \\
&k < \frac{f(\boldsymbol{w}_0) - f(\boldsymbol{w}_k)}{f(\boldsymbol{w}_{k-1}) - f(\boldsymbol{w}_k)}.
\end{aligned}
\tag{18}
$$

Therefore, the proof is complete.

C. Proof of Lemma 2

This proof is similar to A-B, when $k_0 = k_f$, but with considering adaptive b_k. Since at each iteration k, we compute $b_{k+1} = \lceil \eta_k b_k \rceil$, the possible causal way to obtain K is to use the current information of communication energy E_k/b_k. Thus, we obtain K when the causal approximation of E_{k+1},

i.e., $b_{k+1}E_k/b_k$ is greater than $E - \sum_{k'=1}^{k} E_{k'}$. Thus, we obtain the inequality (13).

D. Proof of Proposition 2

According to [16],

$$\begin{aligned}\|\varepsilon_{k+1}^j\|_\infty^2 &\le \tau^2 (R_{k+1}^j)^2 \\ &\le 3\tau^2 L_j \|\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\|_2^2 + 3\tau^2 \|\varepsilon_k^j\|_\infty^2,\end{aligned} \tag{19}$$

where $\|\varepsilon_k^j\|_\infty^2 \le \tau^2 (R_k^j)^2$,

$$\tau^2 (R_{k+1}^j)^2 \le 3\tau^2 L_j \|\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\|_2^2 + 3\tau^4 (R_k^j)^2. \tag{20}$$

According to (19), we derive the following inequality for A-LAQ.

$$\tau_{k+1}^2 (R_{k+1}^j)^2 \le 3\tau_{k+1}^2 L_j \|\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\|_2^2 + 3\tau_{k+1}^2 \tau_k^2 (R_k^j)^2. \tag{21}$$

By inserting (21) into (16) , we obtain the one-step Lyapunov function as

$$\begin{aligned}\mathbb{V}(\boldsymbol{w}_{k+1}) - \mathbb{V}(\boldsymbol{w}_k) \le & -\alpha \langle \nabla f(\boldsymbol{w}_k), \boldsymbol{q}_k \rangle + \frac{\alpha}{2} \|\nabla f(\boldsymbol{w}_k)\|_2^2 \\ & + (\frac{L}{2} + \beta_1 + 3\gamma\tau_{k+1}^2 L_j^2) \|\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\|_2^2 \\ & + \sum_{i=1}^{k_1 - 1} (\beta_{i+1} - \beta_i) \|\boldsymbol{w}_{k+1-i} - \boldsymbol{w}_{k-i}\|_2^2 \\ & - \beta_{k_1} \|\boldsymbol{w}_{k+1-k_1} - \boldsymbol{w}_{k-k_1}\|_2^2 \\ & + \gamma (3\tau_{k+1}^2 - 1) \sum_{j=1}^{M} \|\varepsilon_k^j\|_\infty^2 \\ & + 3\gamma\tau_{k+1}^2 \sum_{j=1}^{M} \|\boldsymbol{q}_{k-1}^j - \boldsymbol{q}_k^j\|_2^2.\end{aligned} \tag{22}$$

By replacing $\boldsymbol{q}_k = \nabla f(\boldsymbol{w}_k) - \varepsilon_k$, $\boldsymbol{w}_{k+1} - \boldsymbol{w}_k = \alpha \boldsymbol{q}_k$, and for any $\rho > 0$

$$\langle \nabla f(\boldsymbol{w}_k), \varepsilon_k \rangle \le \frac{\rho}{2} \|\nabla f(\boldsymbol{w}_k)\|_2^2 + \frac{1}{2\rho} \|\varepsilon_k\|_2^2, \tag{23}$$

and defining $A_{k+1} := L + 2\beta_1 + 6\gamma\tau_{k+1}^2 L_j^2$, we simplify (22) as

$$\begin{aligned}\mathbb{V}(\boldsymbol{w}_{k+1}) - \mathbb{V}(\boldsymbol{w}_k) \le & \|\nabla f(\boldsymbol{w}_k)\|_2^2 \Big(\alpha^2 A_{k+1} - \frac{\alpha}{2} + \frac{\alpha\rho}{2} \Big) \\ & + \|\varepsilon_k\|_2^2 \left(\alpha^2 A_{k+1} + \frac{\alpha}{2\rho} \right) \\ & + \left(\frac{3\gamma\tau_{k+1}^2 \zeta_{k_1}}{\alpha^2 M} - \beta_{k_1} \right) \|\boldsymbol{w}_{k+1-k_1} - \boldsymbol{w}_{k-k_1}\|_2^2 \\ & + \sum_{i=1}^{k_1 - 1} \left(\beta_{i+1} - \beta_i + \frac{3\gamma\tau_{k+1}^2 \zeta_i}{\alpha^2 M} \right) \|\boldsymbol{w}_{k+1-i} \\ & - \boldsymbol{w}_{k-i}\|_2^2 + \gamma \left(3\tau_{k+1}^2 - 1 \right) \sum_{j=1}^{M} \|\varepsilon_k^j\|_\infty^2 \\ \le & \|\nabla f(\boldsymbol{w}_k)\|_2^2 \Big(\alpha^2 A_{k+1} - \frac{\alpha}{2} + \frac{\alpha\rho}{2} \Big) \\ & + \left(\frac{3\gamma\tau_{k+1}^2 \zeta_{k_1}}{\alpha^2 M} - \beta_{k_1} \right) \|\boldsymbol{w}_{k+1-k_1} - \boldsymbol{w}_{k-k_1}\|_2^2 \\ & + \sum_{i=1}^{k_1 - 1} \left(\beta_{i+1} - \beta_i + \frac{3\gamma\tau_{k+1}^2 \zeta_i}{\alpha^2 M} \right) \|\boldsymbol{w}_{k+1-i} - \\ & \boldsymbol{w}_{k-i}\|_2^2 + \left(d\alpha^2 A_{k+1} + \frac{d\alpha}{2\rho} + \gamma \left(3\tau_{k+1}^2 - 1 \right) \right) \times \\ & \left[\sum_{j=1}^{M} \|\varepsilon_k^j\|_\infty \right]^2.\end{aligned} \tag{24}$$

Then, by setting the coefficient to be non-positive, we complete the proof.