Communication Efficient Federated Learning with Adaptive Quantization

Yuzhu Mao[1], Zihao Zhao[1], Guangfeng Yan[2], Yang Liu[3], Tian Lan[4], Linqi Song[2], Wenbo Ding[1]

[1] Tsinghua-Berkeley Shenzhen Institute, [2] City University of Hong Kong, [3] Webank, [4] George Washington University

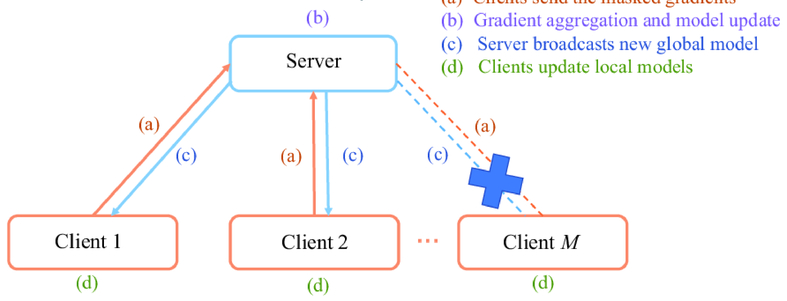
ABSTRACT

We propose 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, we take the client dropout issues into account and develop the Augmented AQG which could limit the dropout noise with amplification for transmitted gradients. Experimental results show that the proposed AQG leads to significant transmission reduction with heterogenous data distributions as compared to existing popular methods, and stays robust to a client dropping rate up to 90%.

Motivation

Two bottlenecks of federated learning (FL):

- High communication overheads
- Limited device reliability



Notations

g_m^k: gradient computed by client m at iteration k

$\hat{g}_m^k$: 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 $\hat{g}_m^k$

$Q_b(g_m^k)$: g_m^k quantized with b bits

$\varepsilon_b(g_m^k)$: quantization error $Q_b(g_m^k)$- g_m^k

θ^k: the aggregated global model broadcasted at iteration k

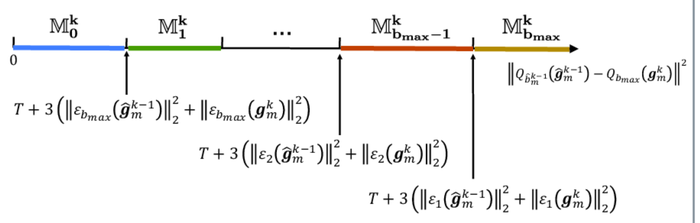
$\mathbb{M}_b^k$: subset of clients uploading gradients with b bits at iteration k

p : client dropping rate

Methods

1. Precision Selection Criterion

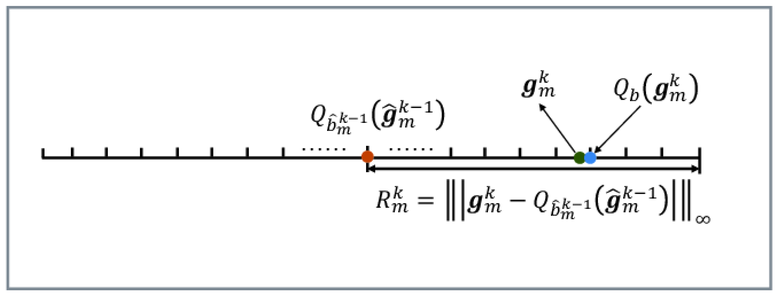
$$T = \frac{1}{\alpha^2 M^2} \sum_{d=1}^{D} \xi_d \left\| \theta^{k+1-d} - \theta^{k-d} \right\|_2^2$$



Under a pre-set upper bound b_{max} for quantization bits:

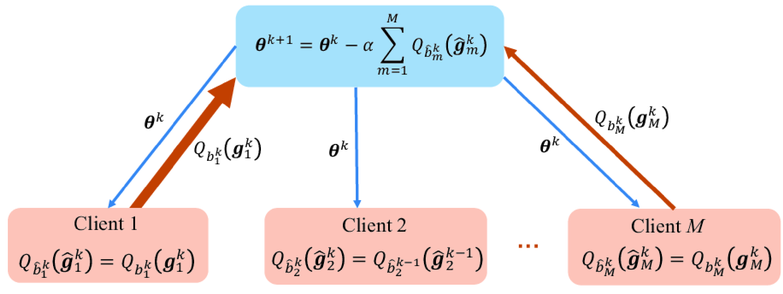
- *Smaller* innovations, *less* number of bits for quantization.
- Too small innovations will be skipped.

2. Quantization Scheme



- Transmission bits for g_m^k : 32/64 bits → b bits

3. Federated Learning with Adaptive Quantized Gradient



4. Augmented AQG with Amplification for Client Dropouts :

$$E\left[Q_{b_m^k}(g_m^k)\right] = (1-p) \cdot \left(\frac{Q_{b_m^k}(g_m^k)}{1-p}\right) + p \cdot \vec{0} = Q_{b_m^k}(g_m^k)$$

Experimental Results

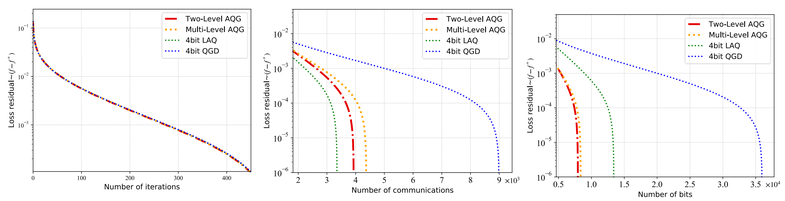


Fig. 1. Convergence of loss function with logistic regression and IID data distribution.

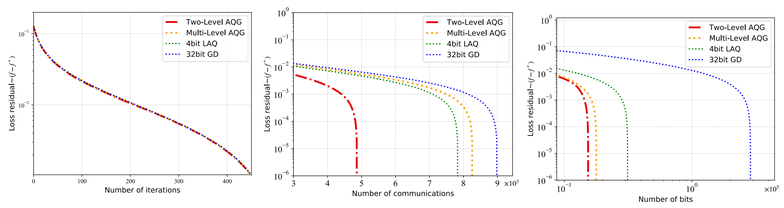


Fig. 2. Convergence of loss function with logistic regression and non-IID data distribution.

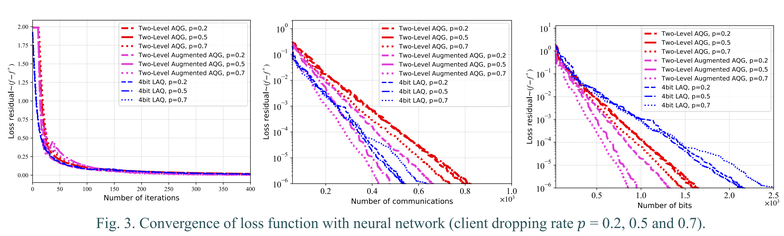


Fig. 3. Convergence of loss function with neural network (client dropping rate p = 0.2, 0.5 and 0.7).

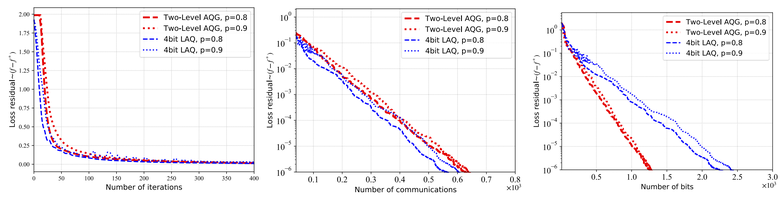


Fig. 4. Convergence of loss function with neural network (client dropping rate p = 0.8 and 0.9).

Conclusions

- AQG leads to 25%-50% of additional transmission reduction compared against existing popular methods.
- AQG with heterogenous data distributions corroborate a more significant transmission reduction compared with independent identical data distributions.
- AQG is robust to a client dropping rate up to 90% empirically.
- The Augmented AQG further improves the FL system's communication efficiency with the presence of moderate-scale client dropouts commonly seen in practical scenes.