Communication-Efficient Federated Learning via Clipped Uniform Quantization

Zavareh Bozorgasl Hao Chen

Department of Eelectrical and Computer Engineering, Boise State University, USA

CISS 2025, 59th Annual Conference on Information Science and Systems
Johns Hopkins University, Baltimore, Maryland
March 19, 2025

Outline

Federated Learning Overview

- Server initializes a global model and sends it to clients.

Federated Learning Overview

- Server initializes a global model and sends it to clients.
- Clients perform local training (e.g., using SGD) on private data.

Federated Learning Overview

- Server initializes a global model and sends it to clients.
- Clients perform local training (e.g., using SGD) on private data.
- Clients send model updates (quantized weights and scaling factors) back to the server.

Federated Learning Overview

- Server initializes a global model and sends it to clients.
- Clients perform local training (e.g., using SGD) on private data.
- Clients send model updates (quantized weights and scaling factors) back to the server.
- The server aggregates the updates (using FedAvg[1] or error-weighted averaging) and redistributes the updated model.

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artificial Intelligence and Statistics (AISTATS)*, PMLR, 2017, pp. 1273–1282.

Federated Learning Overview

- **Federated Learning (FL):** A decentralized training paradigm that preserves data privacy by keeping raw data local.

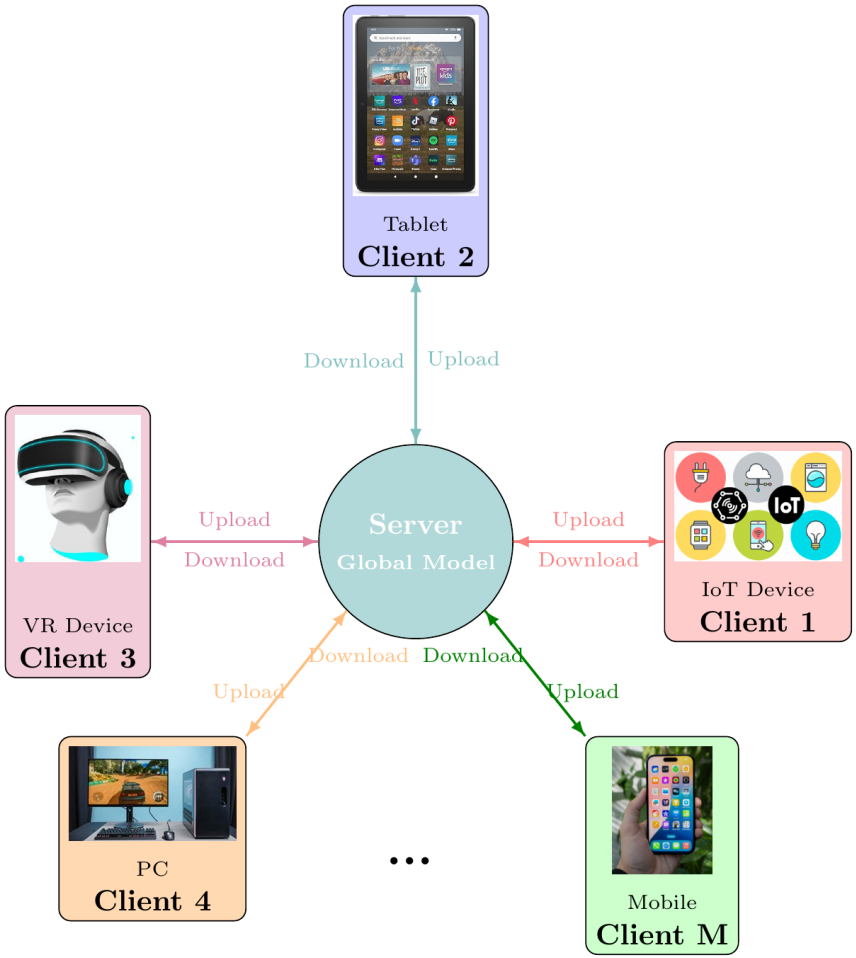

Figure: An illustration of a typical federated learning system.

Our Major Focus: Communication Bottleneck; A Different Weighting

- **Communication Bottleneck:** Transmitting full-precision weight updates is costly.
- **Our Approach:** Use clipped uniform quantization [2]; weighting based on average quantization errors
- **Benefits:**
  - Reduced communication load.
  - Enhanced privacy.
  - Near full-precision performance with resource savings.

[2] R. M. Gray and D. L. Neuhoff, "Quantization," IEEE transactions on information theory, vol. 44, no. 6, pp. 2325–2383, 1998.

Proposed Federated Learning Algorithm

- **Global Model Distribution:** Server sends full-precision model to clients.
- **Local Training:** Clients perform quantization-aware training:
  - Compute optimal clipping thresholds.
  - Clip, quantize, and dequantize weights.
- **Uplink Transmission:** Clients send quantized (K-bit) weights and scaling factors.
- **Aggregation:** Server aggregates updates using FedAvg or error-weighted averaging.
- **Model Update:** Server sends aggregated model back to clients.

Algorithm Outline

Algorithm Steps:

1. **Initialization:** Server distributes global model.
2. **Local Training:** For each round $t = 1$ to T:
   - Compute optimal clipping scalars per layer.
   - Perform clipping, quantization, and dequantization.
3. **Uplink:** Clients send quantized weights and scaling factors.
4. **Aggregation:** Server aggregates updates.
5. **Broadcast:** Server sends updated model to clients.

Clipped Quantization

Why do not we clip outliers?

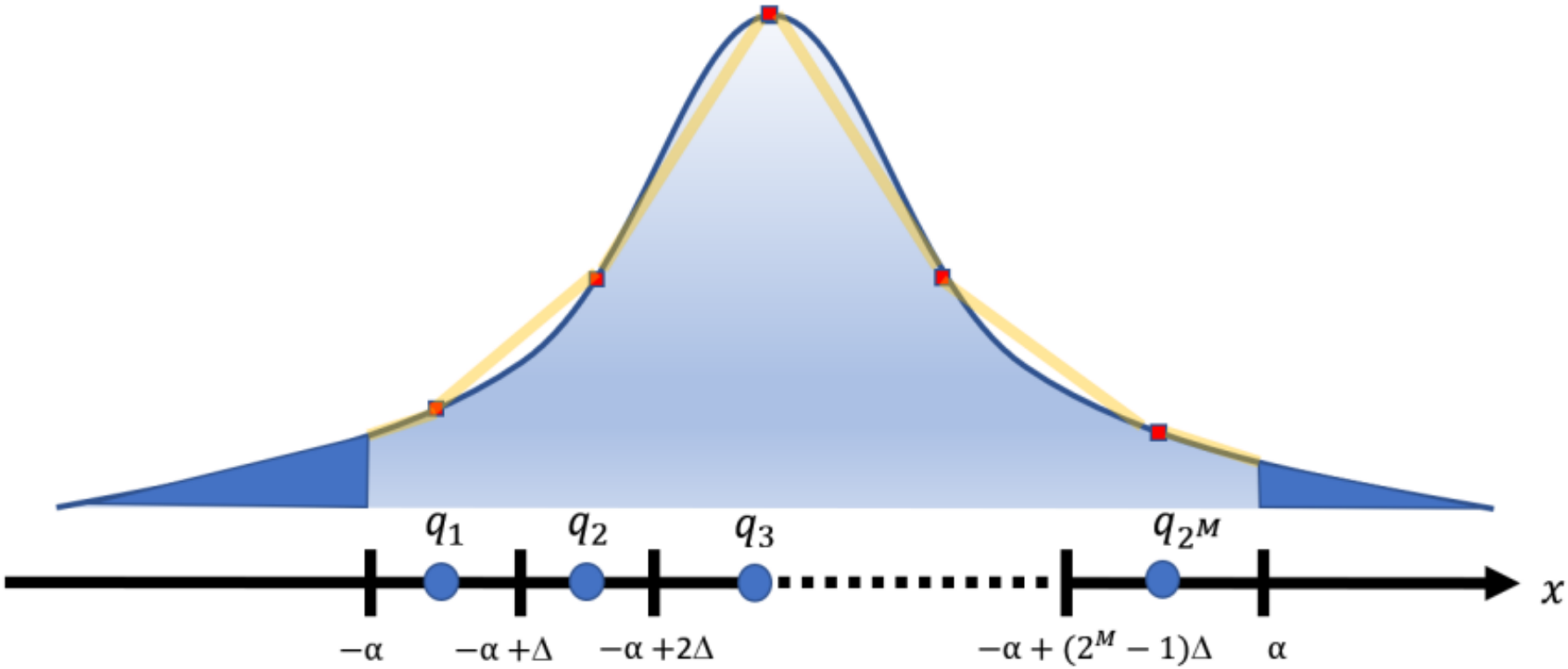

Figure: Uniform Quantization [3]

[3] R. Banner, Y. Nahshan, E. Hoffer, and D. Soudry, "ACIQ: Analytical clipping for integer quantization of neural networks," *arXiv preprint arXiv:1810.05723*, 2018.

Clipped Quantization-ADC

Analog to Digital Converters [4]

- Measurement applications involving temperature, pressure, and weight sensing
- 16-bit

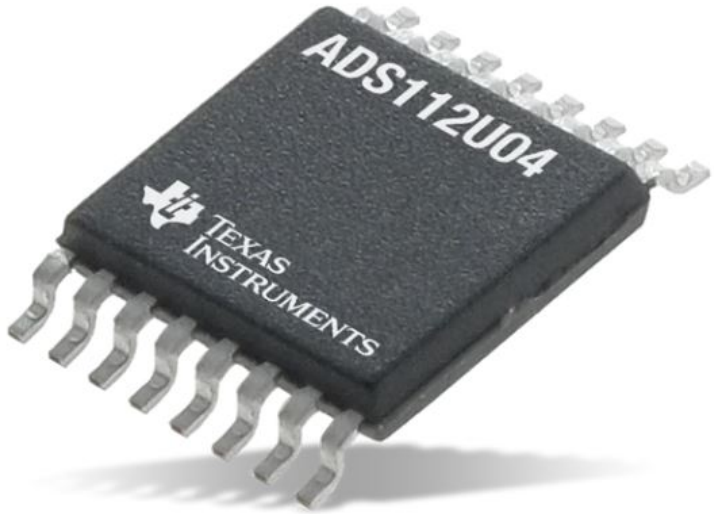

Figure: Application in ADC [5]

[4] N. Al-Dhahir and J. M. Cioffi, "On the uniform ADC bit precision and clip level computation for a gaussian signal," IEEE transactions on signal processing, vol. 44, no. 2, pp. 434–438, 1996.

[5] https://www.ti.com/

Clipped Uniform Quantizers

- **Objective:** Minimize the Mean Squared Error (MSE) between original and quantized values.
- **Clipping:** Restricts dynamic range by mapping values outside $[-s, s]$ to the boundaries.
- **Quantization:** The clipped values are uniformly quantized into $L = 2^b$ levels with step size $\Delta = \frac{2s}{2^b}$.
- **Schemes:**
  - **Deterministic:** Fixed mapping.
  - **Stochastic:** Adds uniform dither to reduce bias.
- An online recursive method is used to determine the optimal clipping threshold s.

Example: Deterministic vs. Stochastic Quantization

Given: Quantizing scalar value $x = 0.8$ using a 2-bit uniform quantizer with levels: $\{-1, -0.33, 0.33, 1\}$

Deterministic Quantization:

- Map x to nearest level: $Q(0.8) = 1$
- Quantization error: $e_{det} = 0.8 - 1 = -0.2$

Stochastic Quantization (probabilistic):

- Probability distribution based on distance:

$$P(q = 0.33) = \frac{1 - 0.8}{0.66} \approx 0.30, \quad P(q = 1) = \frac{0.8 - 0.33}{0.66} \approx 0.70$$

- Expected value:

$$\mathbb{E}[Q(x)] = (0.30 \times 0.33) + (0.70 \times 1) = 0.799$$

- Quantization error:

$$e_{stoch} = 0.8 - 0.799 = 0.001$$

Clipped Quantization- Clipping Value

MSE Expectation Formula

$$E\left[(X - Q(X))^2\right]$$

Threshold Update Formula[6]

The optimal clipping threshold is computed iteratively as:

$$s_{n+1} = \frac{\sum_{x\in\Omega} |x| \mathbf{1}_{\{|x|>s_n\}}}{\frac{4^{-b}}{3} \sum_{x\in\Omega} \mathbf{1}_{\{0<|x|\leq s_n\}} + \sum_{x\in\Omega} \mathbf{1}_{\{|x|>s_n\}}}$$

[6] Sakr et al. (2022). Optimal clipping and magnitude-aware differentiation for improved quantization-aware training, *CVPR*, pp. 11206–11215.

Minimizing Mean Squared Error (MSE)

MSE Minimization Approach

The clipping threshold s is optimized by minimizing the Mean Squared Error (MSE), balancing discretization and clipping errors:

$$E\left[(X - Q(X))^2\right] = \int_{-\infty}^{-s} f(x)(x+s)^2 dx + \sum_{i=0}^{2^b-1} \int_{-s+i\Delta}^{-s+(i+1)\Delta} f(x)(x-q_i)^2 dx + \int_{s}^{\infty} f(x)(x-s)^2 dx$$

Lloyd-Max Derivation for Clipped Quantization

- **Clipped Quantization:**

  To reduce quantization error for non-uniform data, one may clip the data to a range $[-s, s]$ (with $s < s_{\max}$):

$$Q(x) = \begin{cases} -s, & x < -s, \\ s \cdot 2^{1-B} \cdot \text{round}\left(\frac{x}{s} 2^{B-1}\right), & x \in [-s, s], \\ s, & x > s. \end{cases}$$

  The corresponding MSE for [-s,s] becomes

$$J(s) = \frac{4^{-B}}{3} s^2 \int_0^s f_{|X|}(x)\, dx$$

- **Optimal Clipping via Lloyd-Max Approach:**

  The optimal clipping threshold s^* minimizes $J(s)$. Setting $\frac{dJ(s)}{ds} = 0$ and applying the Newton–Raphson method yields an iterative update. This recursion forms the basis of the OCTAV algorithm for dynamically computing optimal clipping scalars.

Aggregation via Quantization Error Weighting

- Weight aggregation based on the inverse of the average squared quantization error.
- For layer i and client j, with dequantized weight $w_{ij,p}$ and average error e_{ij}:

$$\bar{w}_{i,p} = \frac{\sum_{j=1}^{M} \frac{w_{ij,p}}{e_{ij}}}{\sum_{j=1}^{M} \frac{1}{e_{ij}}}.$$

- Clients with lower quantization error contribute more to the aggregated model.

Simulation Setup

- **Model:** Convolutional Neural Network (CNN) for image classification.
- **Datasets:** MNIST and CIFAR-10; IID distribution.
- **FL Setting:** Different number of clients and a global server.
- **Data Size**: Training (50k samples) and test sizes (10k samples)
- **Training:** SGD with fixed hyper-parameters (learning rate 0.01, momentum 0.9, etc.).
- **Configurations:** Various quantization bit-widths (e.g., "4-2-2-4", "8-8-8-8").

Simulation Setup-Configuration for MNIST

Table: Configuration of the neural network employed for the MNIST dataset

Layer Type	**Kernel Size**	**Input Size**	**Output Size**
Conv2d (conv1)	3x3	1 x 28 x 28	16 x 28 x 28
BatchNorm2d (bn1)	N/A	16 x 28 x 28	16 x 28 x 28
MaxPool2d (maxpool1)	2x2	16 x 28 x 28	16 x 14 x 14
Conv2d (conv2)	3x3	16 x 14 x 14	16 x 14 x 14
BatchNorm2d (bn2)	N/A	16 x 14 x 14	16 x 14 x 14
MaxPool2d (maxpool2)	2x2	16 x 14 x 14	16 x 7 x 7
Linear (fc1)	N/A	784 (flattened)	100
BatchNorm1d (bn3)	N/A	100	100
Linear (fc2)	N/A	100	10
BatchNorm1d (bn4)	N/A	10	10
Softmax	N/A	10	10

Communication and Memory Savings (MNIST)

Table: Communication and memory savings for various weight quantization bit-width configurations

Model Weights	FL with Clipping and Quantization of Weights	FL with Full Precision Weights	Communication Saving Times
4-4-4-4	$80848 \times 4 + 4 \times 32$	80848×32	≈ 8
4-2-2-4	$144 \times 4 + 2304 \times 4 + 78400 \times 2 + 1000 \times 4 + 4 \times 32$	80848×32	15.53
2-2-2-2	$144 \times 2 + 2304 \times 2 + 78400 \times 2 + 1000 \times 2 + 4 \times 32$	80848×32	15.98
2-1-1-2	$144 \times 2 + 2304 \times 1 + 78400 \times 1 + 1000 \times 2 + 4 \times 32$	80848×32	31.12

Histograms of weights of NN Models

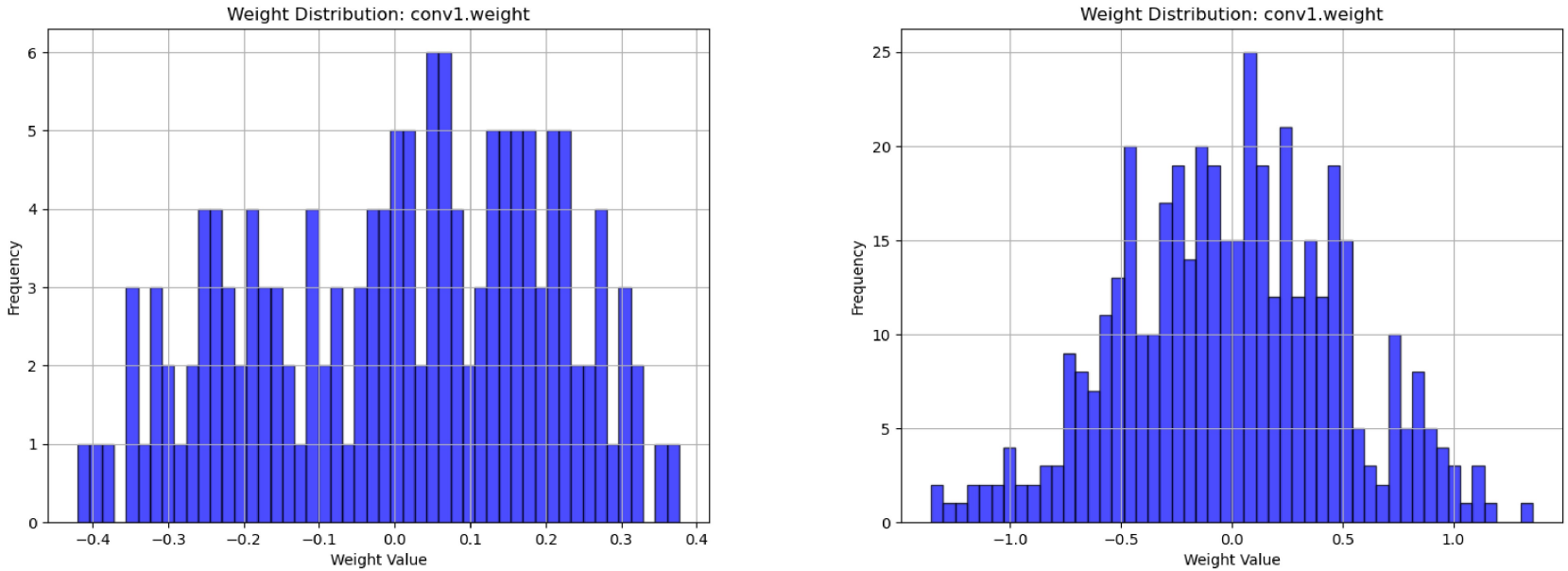

Figure: Comparison of histogram of weights of the first layers of configurations on the MNIST (left) and CIFAR-10 (right) datasets after 100 and 500 rounds of training, respectively.

Simulation Results (MNIST)-"4-2-2-4" Config.

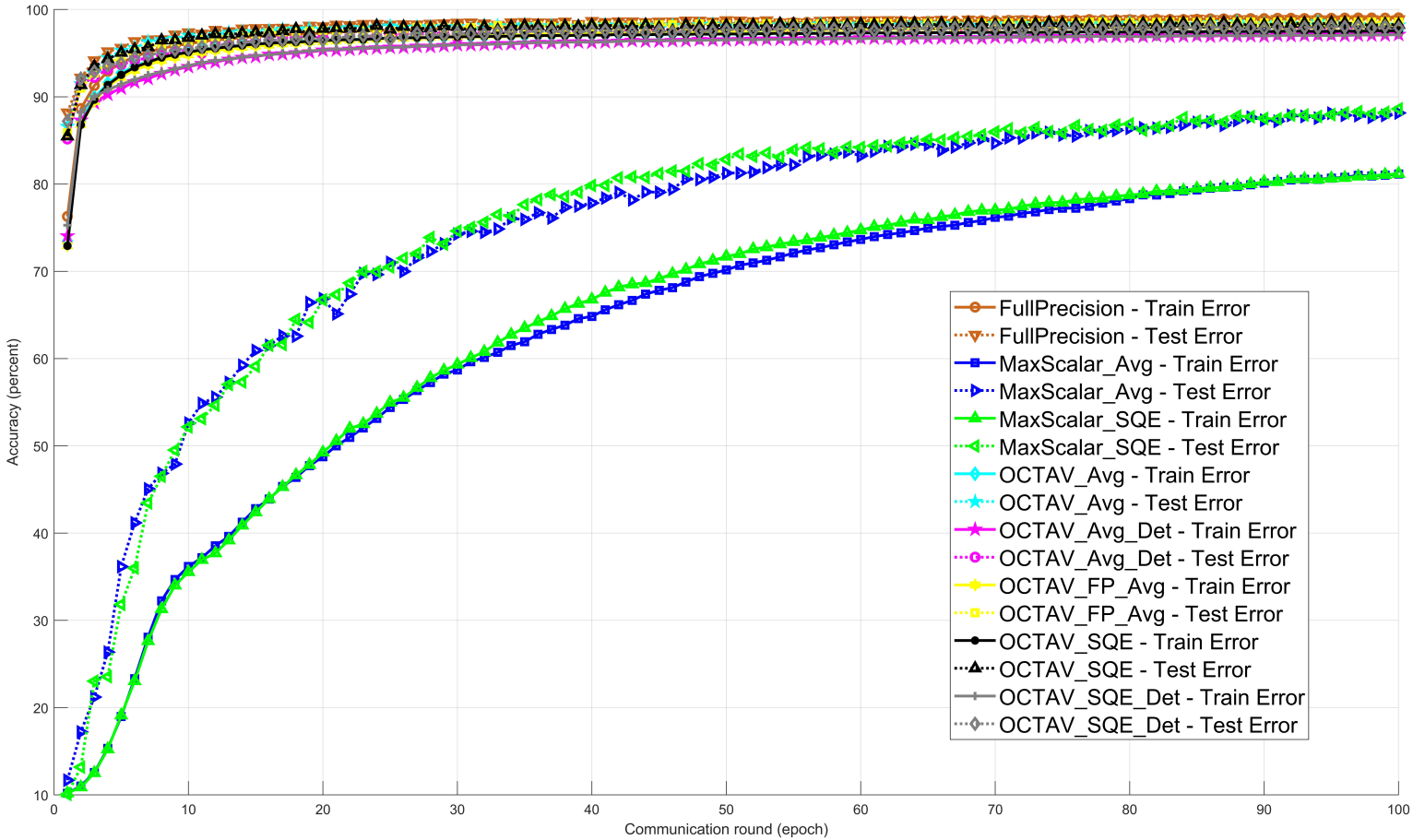

Figure: Communication rounds (Epoch) versus training and test accuracy (percent) for MNIST dataset ("4-2-2-4").

Simulation Results (MNIST)-Different no. of Clients

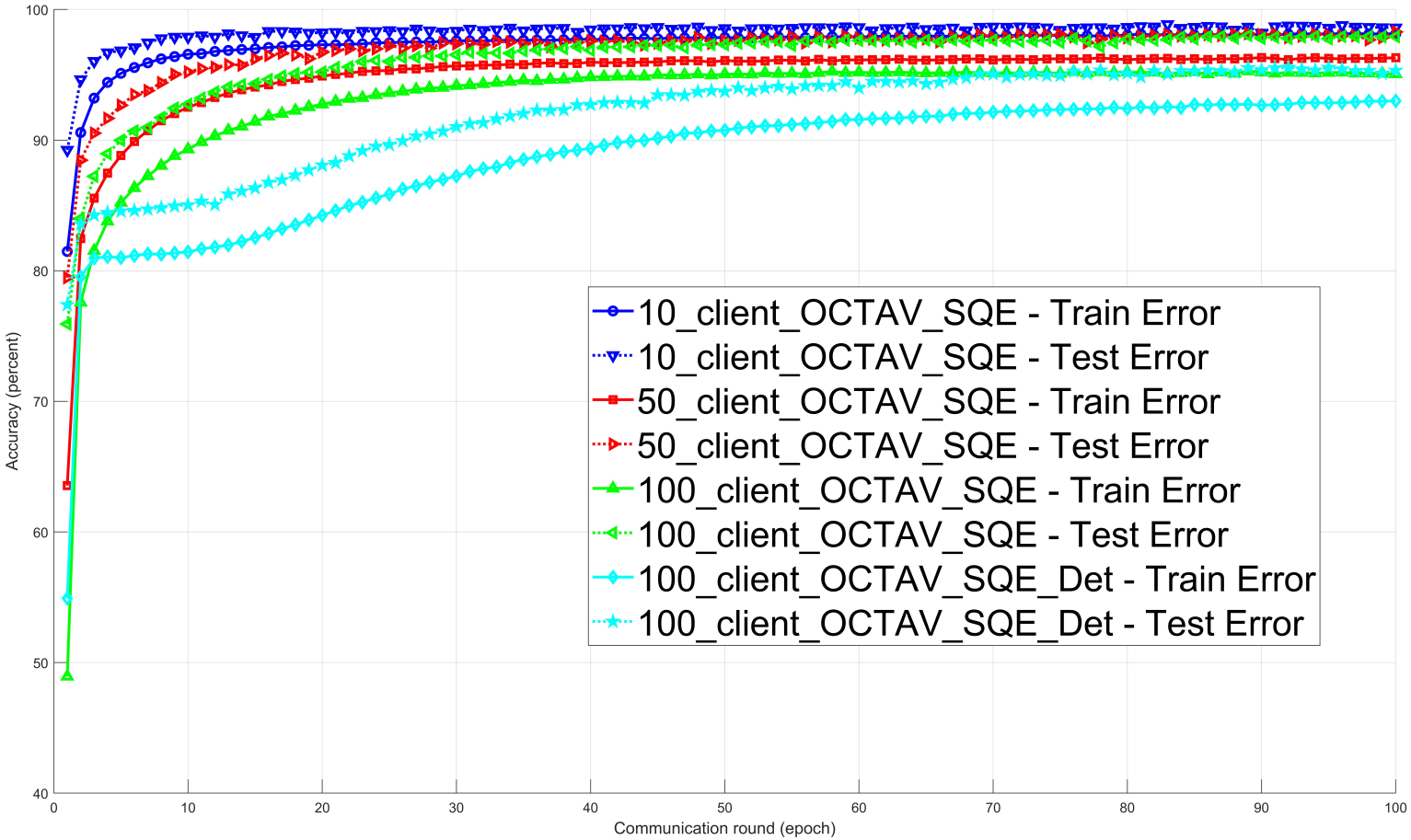

Figure: Communication rounds (Epoch) versus training and test accuracy (percent) for MNIST dataset for different number of clients ("2-2-2-2").

Simulation Results (CIFAR-10)-"4-2-2-4" Config.

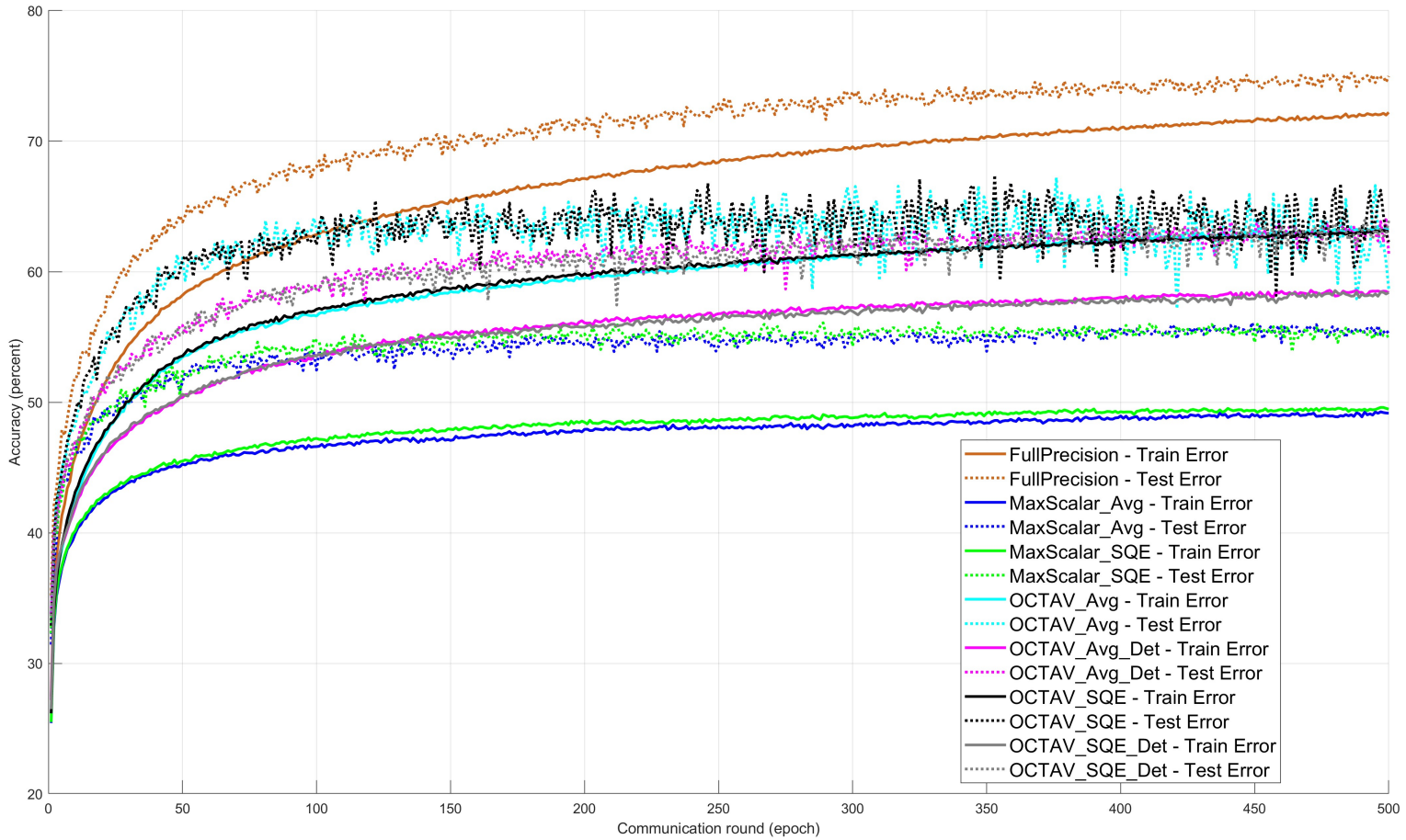

Figure: Communication rounds (Epoch) versus training and test accuracy (percent) for CIFAR10 dataset ("4-2-2-4").

Simulation Results (CIFAR-10)-Different Configurations

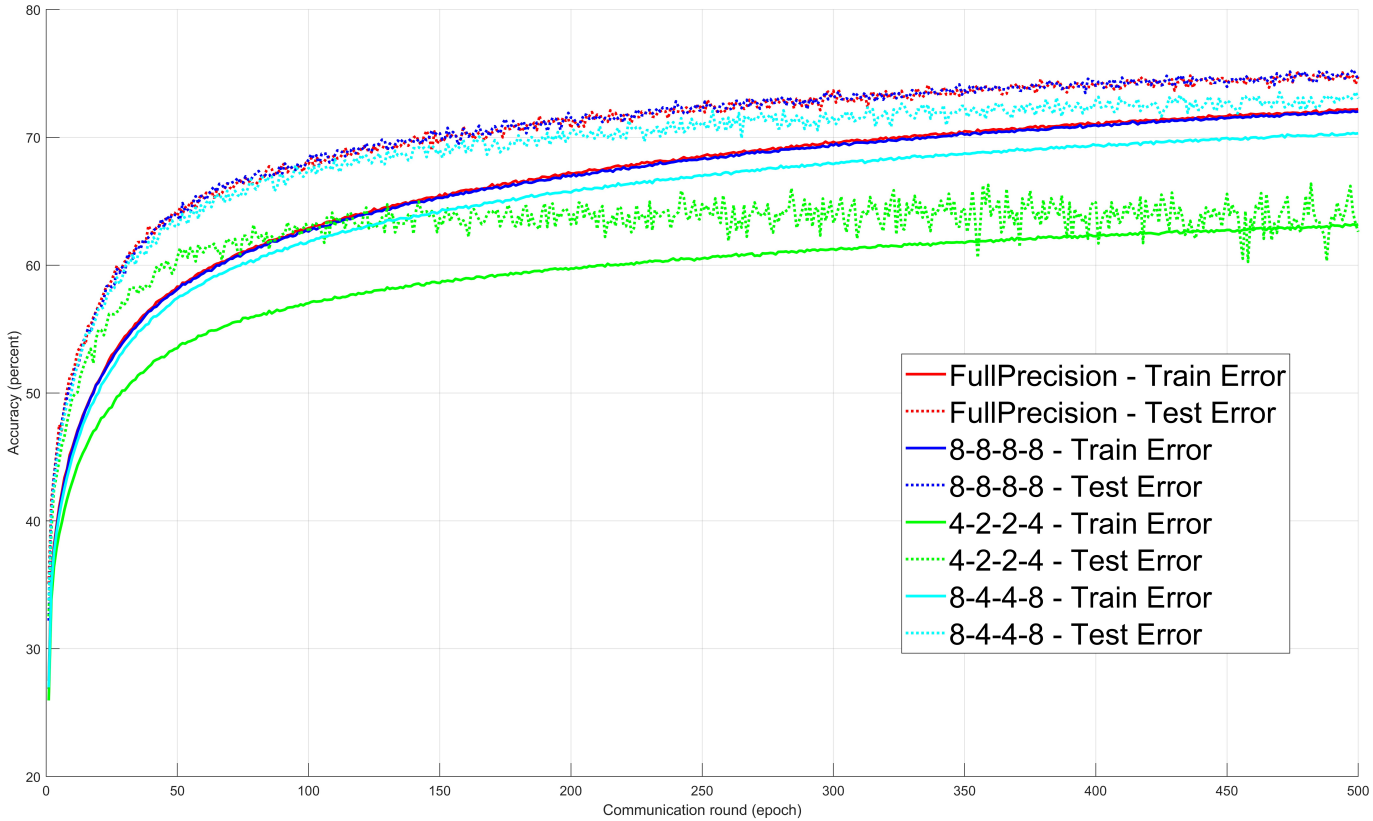

Figure: Comparison of training and test accuracy (percentage) across communication rounds (epochs) for different neural network configurations on the CIFAR-10 dataset.

Additional Analysis

- Clipped quantization (OCTAV) achieves near full-precision accuracy.
- Both FedAvg-based and error-weighted aggregation yield comparable performance.
- For MNIST, less bit in first/layers does not hurt much.

Additional Analysis and Future works

- Histograms indicate higher dynamic range in the first and last layers.
- Per-channel quantization may further improve precision at the cost of increased computation.
- Stochastic quantization demonstrates higher robustness.
- Considering Expectation-Maximization (EM) to refine the clipped weights, replacing them with EM-derived values or a hybrid approach that combines Mean Squared Error (MSE) and EM for enhanced accuracy.

Conclusion

- Introduced a framework for communication-efficient federated learning using clipped uniform quantization.
- Leveraged optimal clipping thresholds with both deterministic and stochastic quantization.
- Explored two aggregation strategies: conventional FedAvg and error-weighted averaging.
- Simulations on MNIST and CIFAR-10 demonstrate significant communication/memory savings with near full-precision performance.

Codes and Slides

Email: Zavarehbozorgasl@u.boisestate.edu

`github.com/zavareh1/ClippedQuantFL`

Acknowledgments

- We thank our colleagues, especially Dr. Aykut Satici and Dr. John Chiasson.
- Dr. John Chiasson was an IEEE fellow who passed away on February 27, 2025. RIP John!

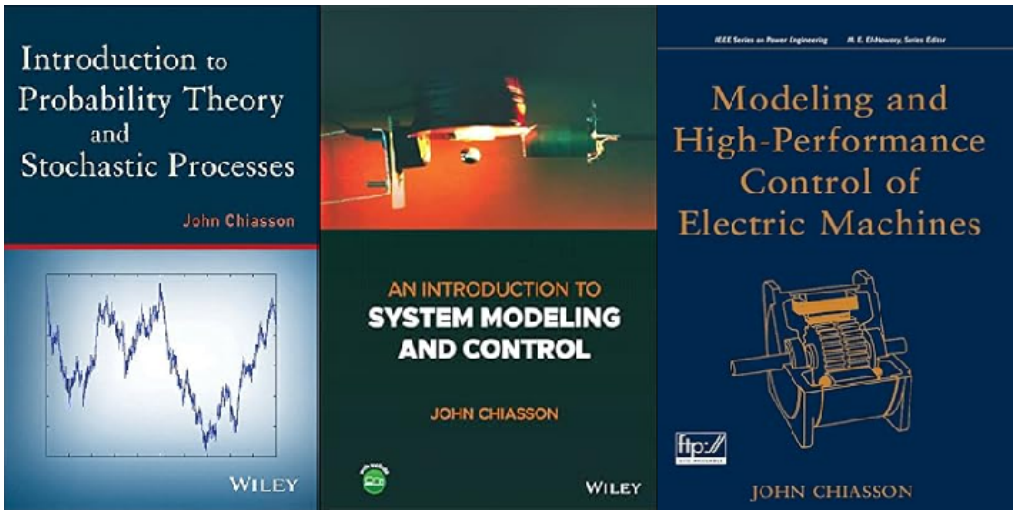

Figure: Some of his books.

Q&A

Thank You!

Questions?