

SQMG: An Optimized Stochastic Quantization Method Using Multivariate Gaussians for Distributed Learning

**Jianan Zhang^{*‡}, Zaipeng Xie^{*‡}, Hongxing Li[†], Xuanyao Jie[‡], Yunfei Wang[‡],
and Bowen Li^{*‡}**

^{}Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University*

[†]Dayu College, Hohai University

[‡]Department of Computer Science and Technology, Hohai University, China



Outline

- Background
- Motivation
- Method
- Experiments
- Conclusion

Background - Distributed Learning

■ Distributed Learning

- ❑ Training models across multiple computational nodes, leveraging parallel processing for large datasets and complex models.

■ Core Challenges^[1]

- ❑ **Communication Overhead:** High costs in synchronizing models across nodes.
- ❑ **Model Synchronization:** Faster nodes wait for slower ones during model synchronization.
- ❑ **Balancing efficiency and performance:** Trade-off where programs designed to improve efficiency can unintentionally amplify existing biases within the models.

■ Prominent Approach: Asynchronous Stochastic Gradient Descent and Quantization^[2]

[1] Q. Zhou, S. Guo, Z. Qu, P. Li, L. Li, M. Guo, and K. Wang, “Petrel: Heterogeneity-aware distributed deep learning via hybrid synchronization,” IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 5, pp. 1030–1043, 2020.

[2] C. D. Sa, M. Feldman, C. Ré, and K. Olukotun, “Understanding and optimizing asynchronous low-precision stochastic gradient descent,” in Proceedings of the 44th Annual International Symposium on Computer Architecture, ISCA 2017, Toronto, ON, Canada, June 24-28, 2017. ACM, 2017, pp. 561–574.

Background - Reduce Communication Costs

■ Stochastic Quantization^[3]

- Probabilistically quantizes the data to the two points in the quantization target space closest to the original values.

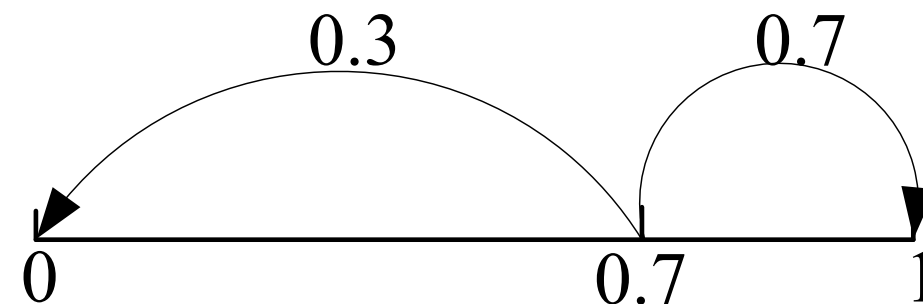
■ Pivotal Elements of Quantization

- Quantization target space
- Quantization mapping scheme

■ Other Methods

- Knowledge Distillation / Model Pruning
- Compressing Communication Messages

Mapping to both ends with different probabilities



$$[0.7, 0.7, \dots, 0.7] \rightarrow [1, 0, \dots, 1]$$

Background - AQSGD

■ Asynchronous Quantized Stochastic Gradient Descent(AQSGD)

- ❑ **Asynchronicity**^[2] allows each node to compute and apply gradients independently, significantly enhancing computational efficiency through minimized idle time.
- ❑ **Stochastic quantization** compresses numerical data like model parameters and gradients, reducing communication and storage needs effectively.

■ Core Challenges ^[4]

- ❑ **Bias in Quantified Spatial Distributions:** Traditional stochastic quantization methods usually assume uniform distributions, which often unintentionally introduces biases.
- ❑ **Balance Between Speed and Accuracy:** Finding the right level of quantization—enough to speed up learning but not so much as to degrade the model's performance—is crucial.

[2] C. D. Sa, M. Feldman, C. Ré, and K. Olukotun, “Understanding and optimizing asynchronous low-precision stochastic gradient descent,” in Proceedings of the 44th Annual International Symposium on Computer Architecture, ISCA 2017, Toronto, ON, Canada, June 24-28, 2017. ACM, 2017, pp. 561–574.

[4] L. Liu, J. Zhang et al., “Hierarchical federated learning with quantization: Convergence analysis and system design,” IEEE Transactions on Wireless Communications, vol. 22, no. 1, pp. 2–18, 2022.

Motivation

■ Enhancing Communication Efficiency

- ❑ There is a critical need for methods that reduce data transmission volume without sacrificing the integrity of the information.

■ Overcoming Conventional Quantization Limitations

- ❑ Conventional quantization techniques typically use a uniform scaling ratio, which can lead to large cumulative quantization errors.

■ Utilizing Data Correlations

- ❑ Existing quantization methods struggle to balance communication efficiency with the accuracy of the model updates—either compromising speed for accuracy or vice versa.

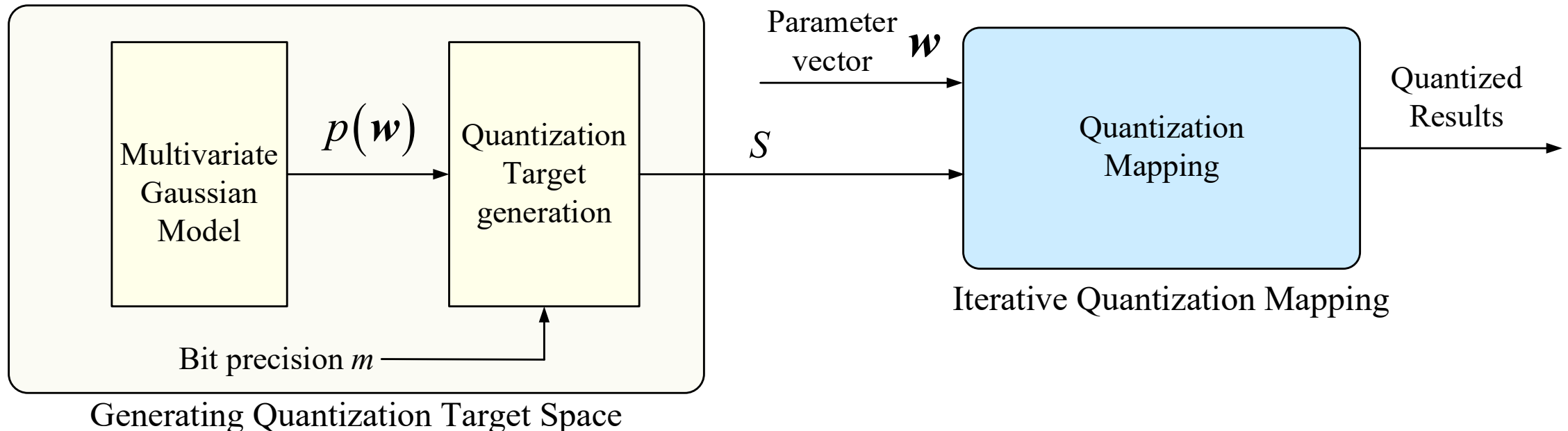
■ Improving Model Convergence and Accuracy

- ❑ Accelerating the convergence of learning models is paramount for reducing computational and operational costs.

Method - Overall

■ SQMG (Stochastic Quantization using Multivariate Gaussians)

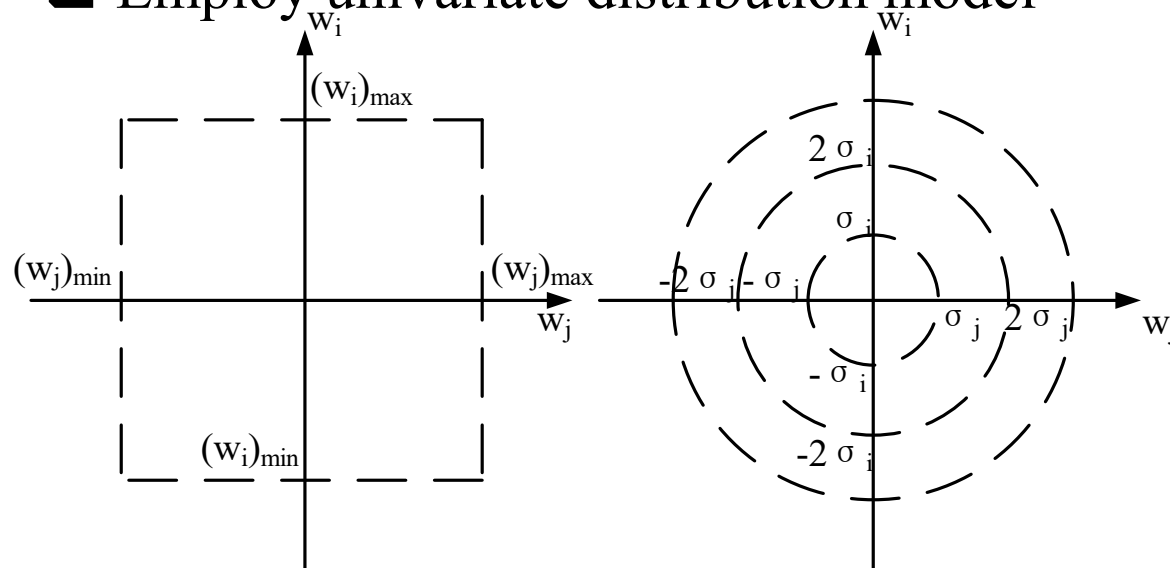
- ❑ **Generating Quantization Target Space:** Models the joint distribution of gradients using a multivariate Gaussian approach, allowing for a more accurate representation of the gradients' relationships and variances.
- ❑ **Iterative Quantization Mapping Scheme:** Projects quantized parameters onto an optimized target space, effectively minimizing quantization errors during each training iteration.



Method - Multivariate Gaussian Model

■ Conventional approach :

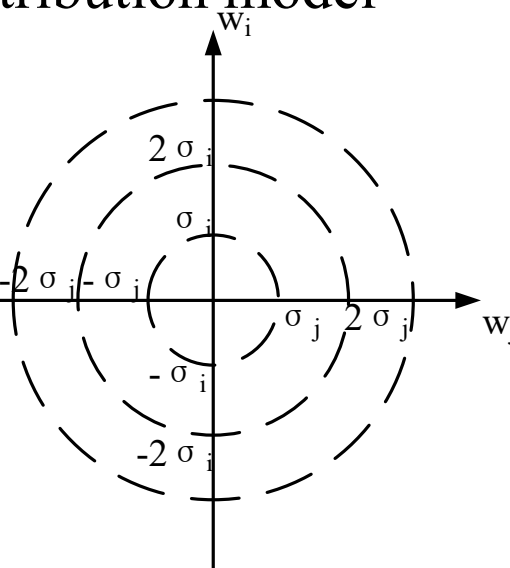
- ❑ Quantize each component of \mathbf{w} individually
- ❑ Assume homogeneous distribution and zero covariance
- ❑ Alter direction and magnitude of quantized vector, leading to significant errors
- ❑ Employ univariate distribution model



Uniform distribution

■ Our Proposed SQMG Approach:

- ❑ Quantize \mathbf{w} holistically using $\alpha \cdot \mathbf{b}$, where $\mathbf{b} \in S$
- ❑ Assume equal probability for gradient direction and nonzero covariance
- ❑ Maintain constant magnitude, vary only direction, reducing errors
- ❑ Employ **multivariate Gaussian model** to represent the entire parameter vector \mathbf{w} using a new **quantization target space**
- ❑ Aim to mitigate errors from non-IID datasets across nodes



Multivariate Gaussian model

Method - Multivariate Gaussian Model

■ Definition 1: Construction of Quantization Target Space \mathcal{S}

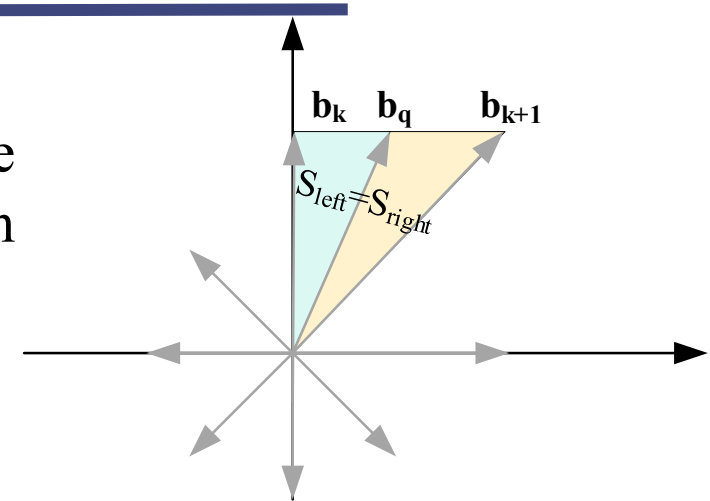
Let $\mathbf{w} \in \mathbb{R}^n$ represent the parameter vector to be quantized. We model the probability distribution of \mathbf{w} as a multivariate Gaussian distribution with mean vector μ and covariance matrix Σ .

$$p(\mathbf{w}) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{w} - \mu)^T \Sigma^{-1}(\mathbf{w} - \mu)\right)$$

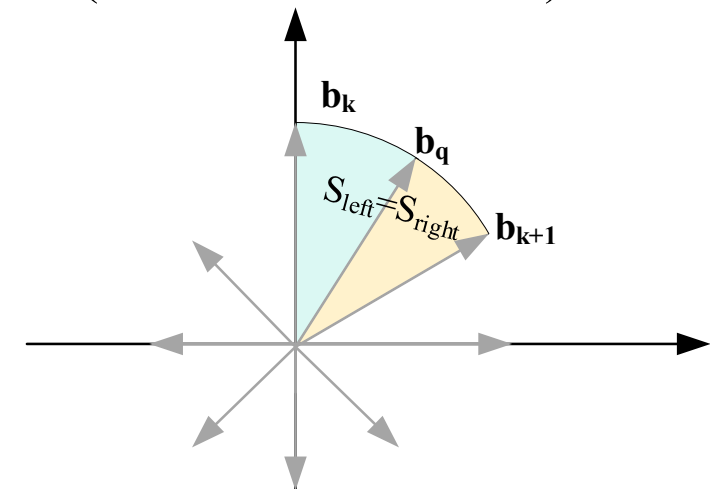
Employing the transformation from polar to Cartesian coordinates, we construct the quantization target space \mathcal{S} under this distributional assumption using the transformation:

$$\mathcal{S} = \left\{ \begin{array}{l} -\theta\left(\frac{1}{k}\right), \dots, -\theta\left(\frac{k-1}{k}\right), \quad -\theta\left(\frac{k}{k}\right), \quad 0, \\ \theta\left(\frac{k}{k}\right), \theta\left(\frac{k-1}{k}\right), \dots, \theta\left(\frac{1}{k}\right) \end{array} \right\},$$

where $\theta(x) = 1/\tan(\pi x/4)$, $k = 2^{m-1} - 1$



Conventional approach
(uniform distribution)



Our SQMG approach
(multivariate Gaussian model)

Method – Iterative Mapping Scheme

■ Once \mathcal{S} is established, the next step is to map the original parameter vector \mathbf{w} to the quantized vector $\alpha \cdot \mathbf{b}$, where $\mathbf{b} \in \mathcal{S}$. The Iterative Mapping Scheme is designed to minimize the quantization error within acceptable quantization error δ .

■ Proposition 1: Quantization Mapping Scheme

The objective function of the iterative quantization mapping can be formulated as follows:

$$\alpha^*, \mathbf{b}^* = \operatorname{argmin}_{\alpha, \mathbf{b}} \|\alpha \cdot \mathbf{b} - \mathbf{w}\|_2^2$$

As a result, the quantization mapping scheme of our proposed SQMG method for the input parameters within the quantized space can be described as:

Algorithm 1 Quantization Mapping Scheme for SQMG

Input: Parameters \mathbf{w} , Acceptable error δ

Output: Quantization coefficient α , quantization base \mathbf{b}

```

1: Initialize  $\alpha, \mathbf{b}_{pre}, \mathbf{b} \leftarrow \delta + 1$ ;
2:  $\mathbf{b} \leftarrow Q\left(\frac{\mathbf{w}}{\alpha}\right)$ ;
3: while  $|\mathbf{b} - \mathbf{b}_{pre}| > \delta$  do
4:    $\alpha \leftarrow \frac{\operatorname{sum}(\mathbf{b} \odot \mathbf{w})}{\operatorname{sum}(\mathbf{b} \odot \mathbf{b})}$ ;
5:    $\mathbf{b}_{pre} \leftarrow \mathbf{b}, \mathbf{b} \leftarrow Q\left(\frac{\mathbf{w}}{\alpha + \varepsilon}\right)$ ;
6: end while
7: return  $\alpha, \mathbf{b}$ 

```

$$\begin{cases} b_i \leftarrow Q\left(\frac{w_i}{\alpha + \varepsilon}\right) \\ \alpha \leftarrow \frac{\|\mathbf{b} \odot \mathbf{w}\|_1}{\|\mathbf{b} \odot \mathbf{b}\|_1} \end{cases},$$

The formal proof of Proposition 1 is detailed on the last page of our paper

Experiments - Setup

	Experimental Settings
Applications	Deep Neural Network (DNN) for MNIST handwritten digit recognition task, Convolutional Neural Network (CNN) for CIFAR-10 image classification task.
Server Settings	Alibaba Cloud’s ECS: 1 parameter server and 10 worker nodes, each equipped with 4GB memory, and Intel Xeon Platinum 8269CY processors at 3.2 GHz
Network Bandwidth	1 Gbps for both upload and download operations
Quantization Accuracy	16-bit floating point (Half precision, HP) for the DNN model baseline 64-bit floating point (Double Precision, DP) for the CNN model baseline 3-bit and 6-bit configurations for all other quantization methods
Evaluation Metric	Accuracy during initial / final training rounds of training/testing set.
Training rounds	3,000 rounds for DNN, 30,000 rounds for CNN

Experiments – Compared Methods

- **Baseline:** *Quantization free*. 16-bit floating point (Half precision, HP) for DNN models, while 64-bit floating point (Double Precision, DP) for CNN models
- **LAQ^[5]**: Considers *quantization's impact on loss* via an approximate Newton algorithm, integrating preprocessing gradient descent and quantization steps. Enables distinct scaling factors for positive and negative weights.
- **SDQ^[6]**: Employs *variable bitwidth* parameters to dynamically determine optimal quantization configurations for each layer. Adjusts precision during training by regulating the likelihood of quantizing activations and weights to layer-specific bitwidths.
- **SQMG** (*Proposed method*): Employs a multivariate Gaussian model to generate quantization target spaces and an iterative mapping algorithm for quantization.

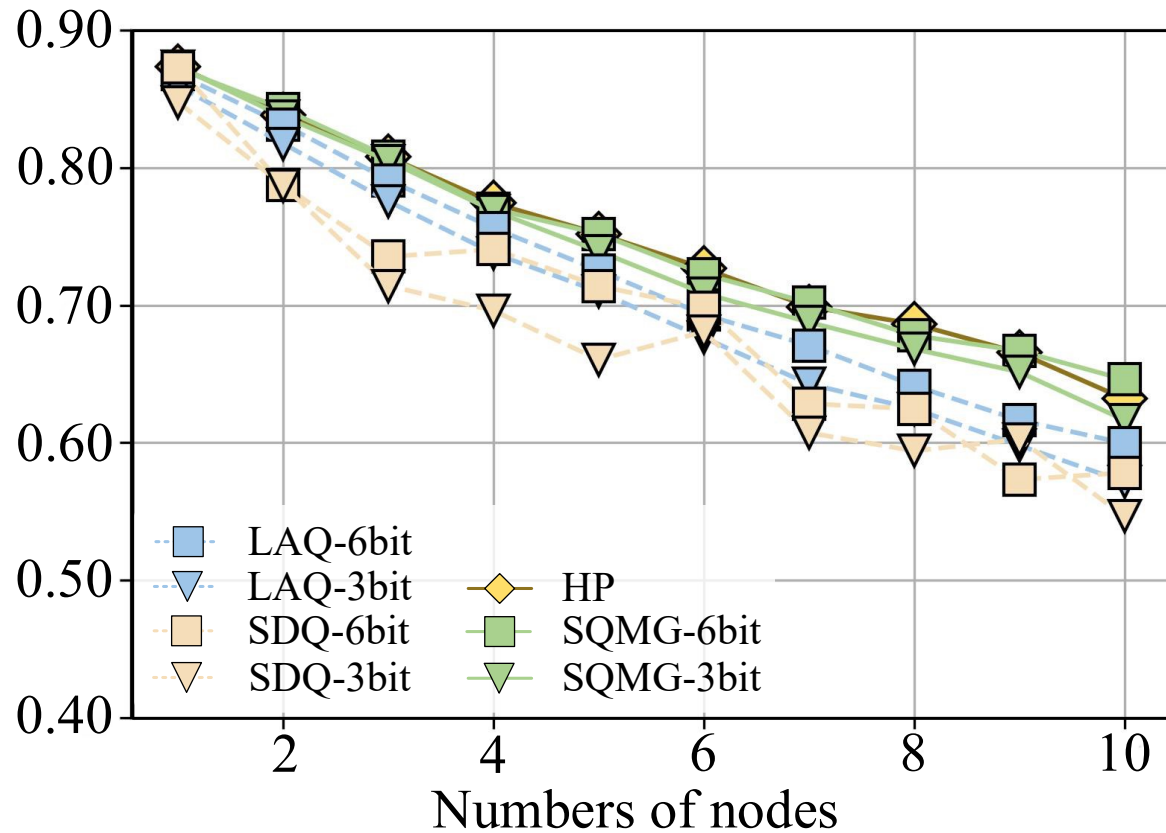
[5] L. Hou, R. Zhang, and J. T. Kwok, “Analysis of quantized models,” in 7th International Conference on Learning Representations, ICLR, New Orleans, LA, USA, May 6-9, 2019.

[6] X. Huang, Z. Shen, S. Li, Z. Liu, H. Xianghong, J. Wicaksana, E. Xing, and K.-T. Cheng, “SDQ: Stochastic differentiable quantization with mixed precision,” in International Conference on Machine Learning, ICML, 2022, pp. 9295–9309. 12

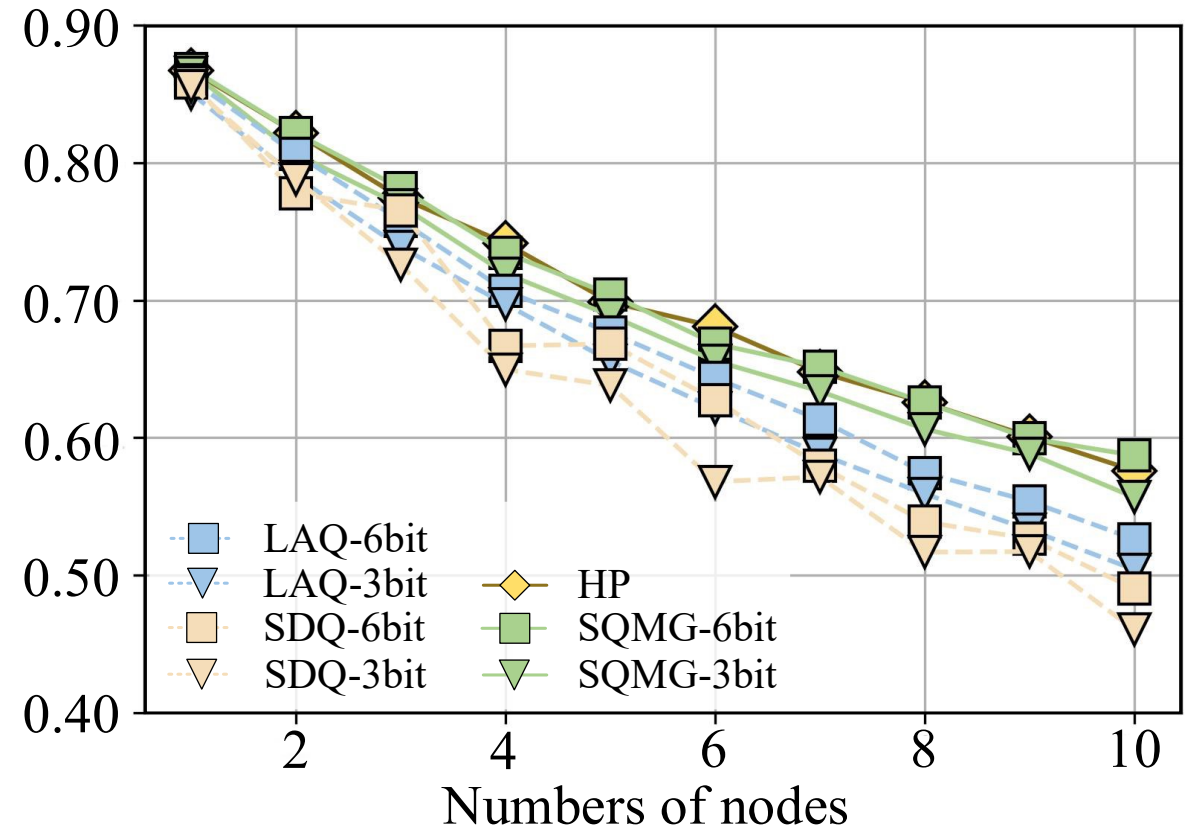
Experimental Results – DNN Model Scenarios

■ Top-1 accuracy of the DNN model after the first training round.

(a) the local training set; (b) the testing set



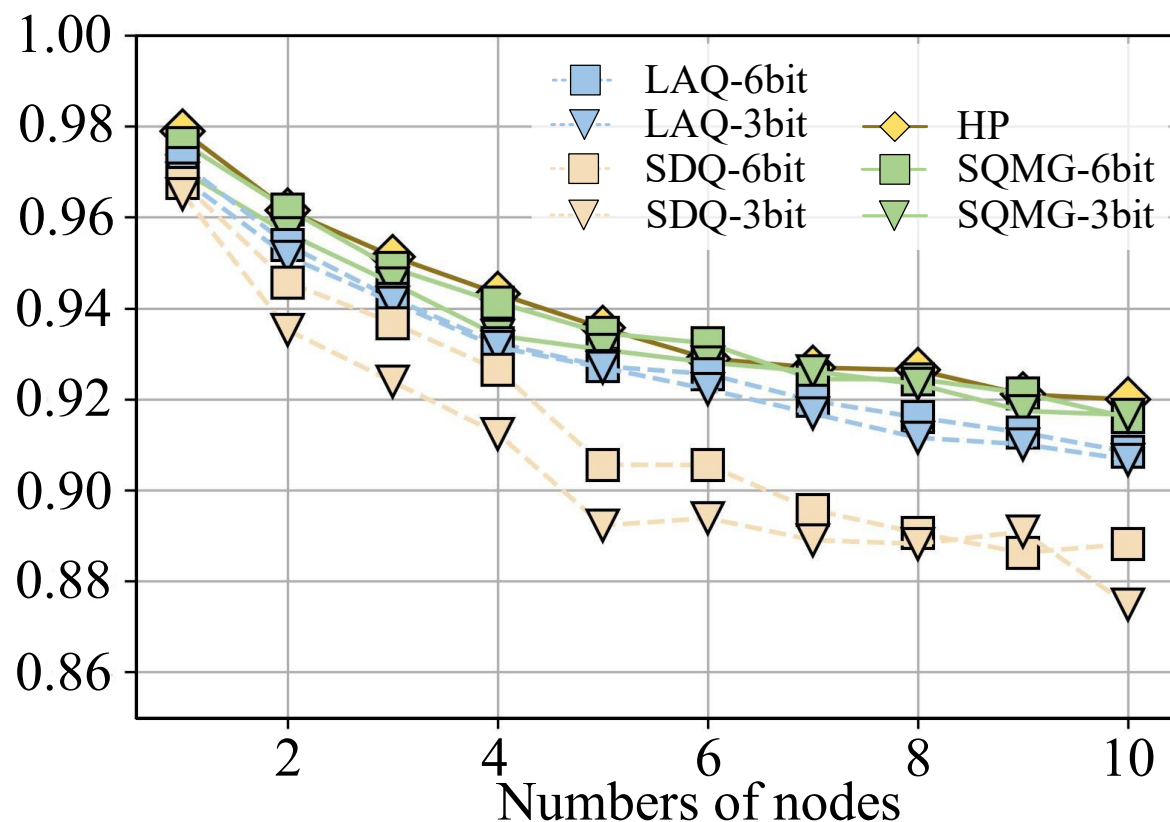
(a) Top-1 Accuracy for local Training Set (First round)



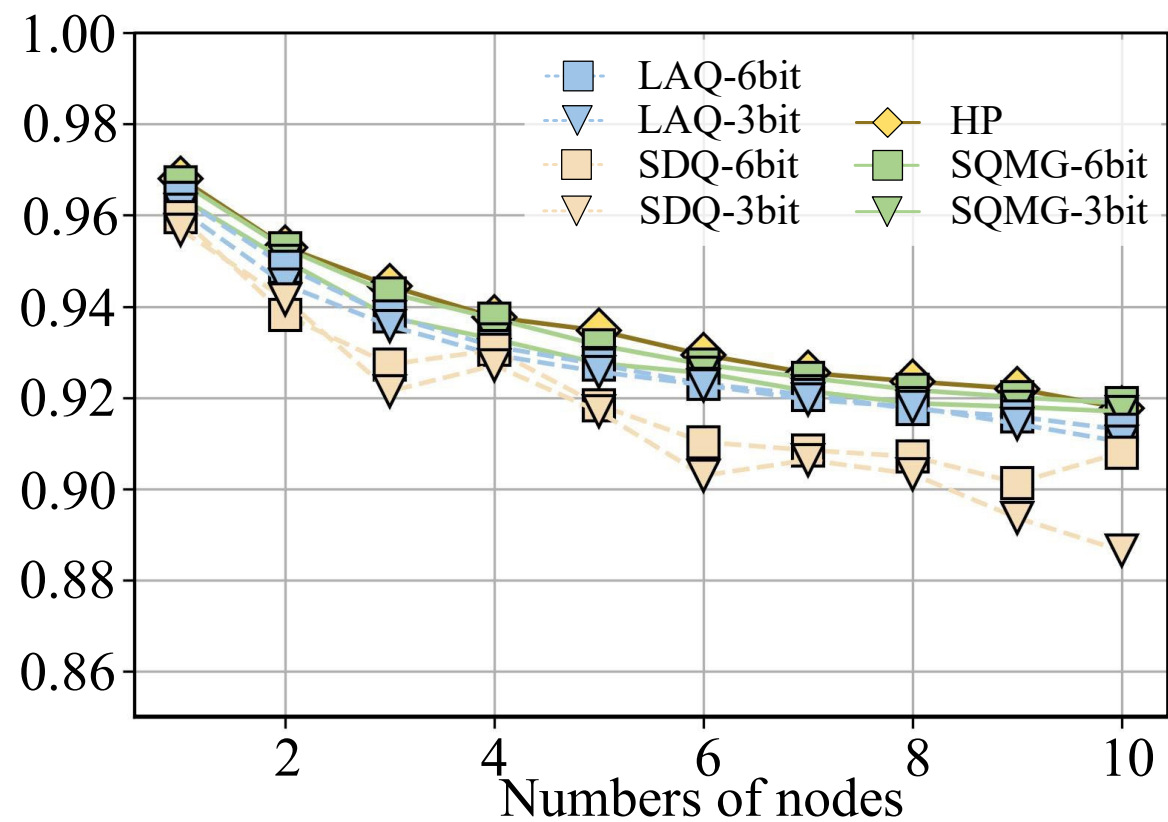
(b) Top-1 Accuracy for Testing Set (First round)

Experimental Results – DNN Model Scenarios

- Top-1 Accuracy of DNN model versus the number of nodes after training is completed.
(a) the local training set; (b) the testing set.



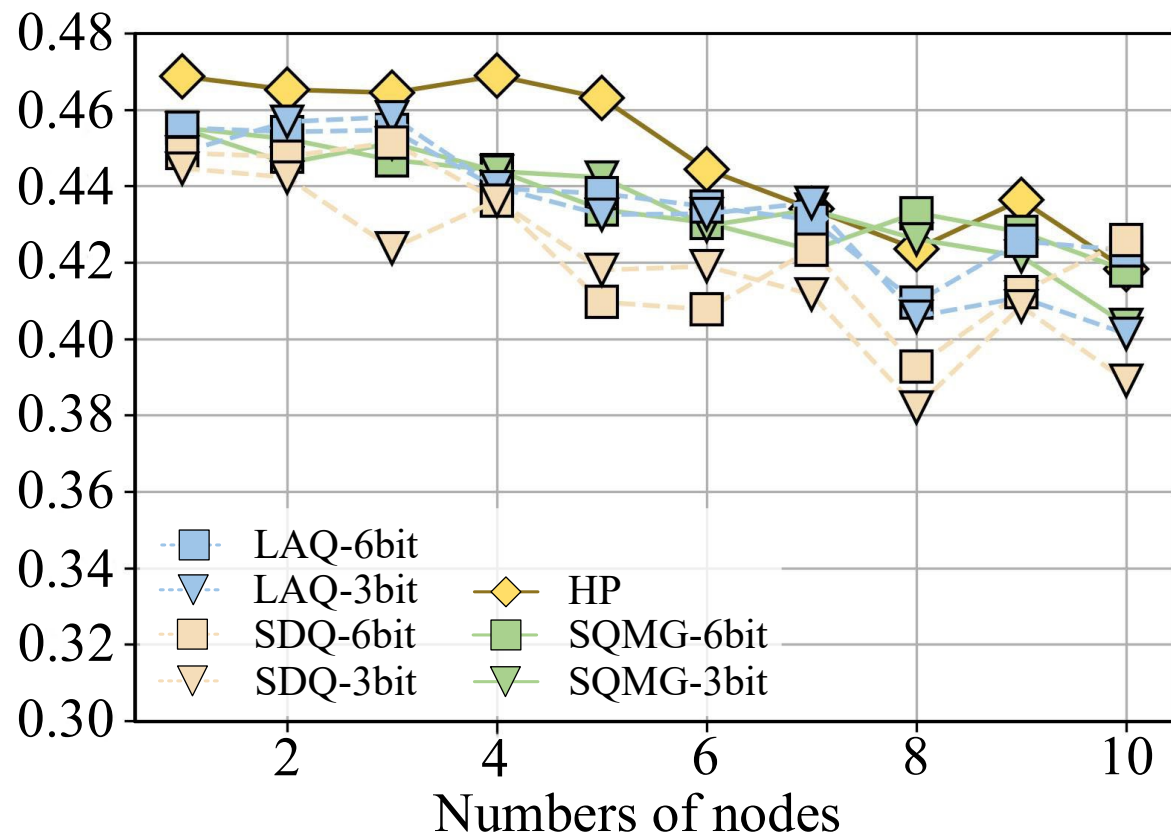
(a) Top-1 Accuracy for local Training Set (Final round)



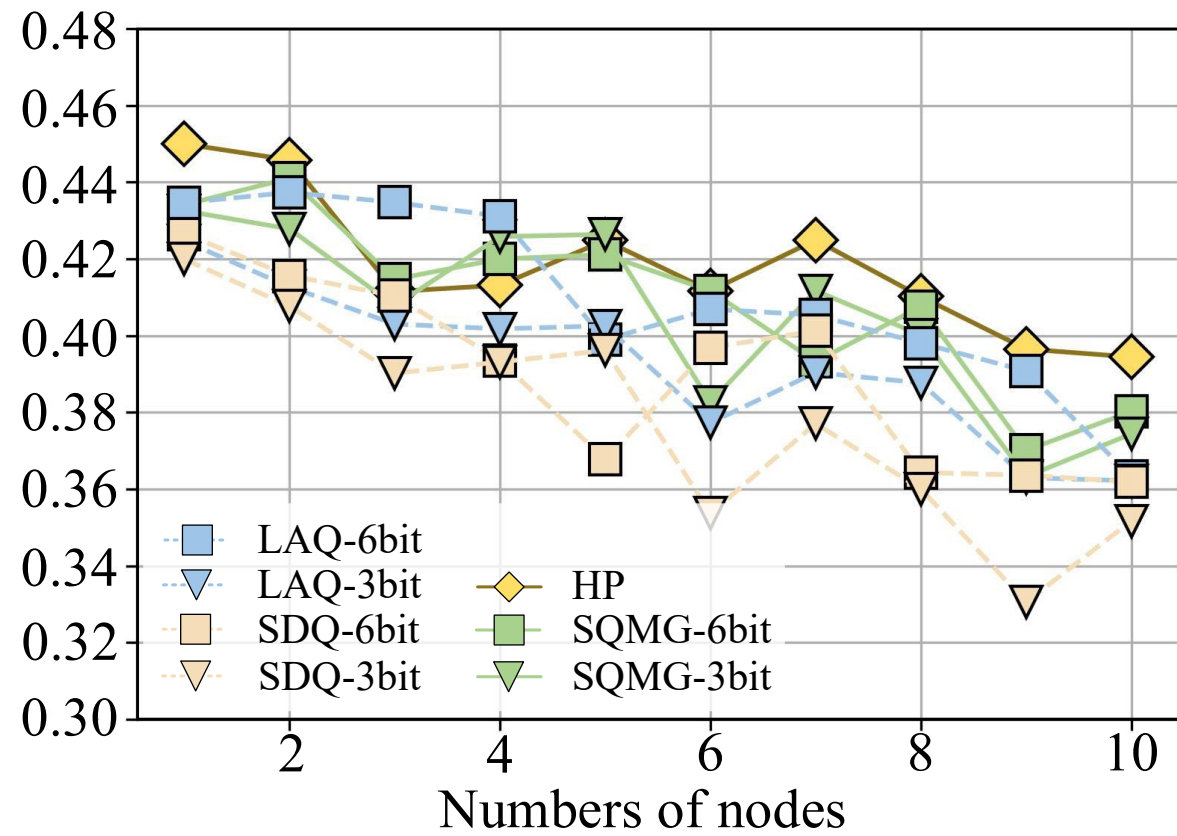
(b) Top-1 Accuracy for Testing Set (Final round)

Experimental Results – CNN Model Scenarios

- Top-1 Accuracy of the CNN model after the first training round.
 (a) the local training set; (b) the testing set.



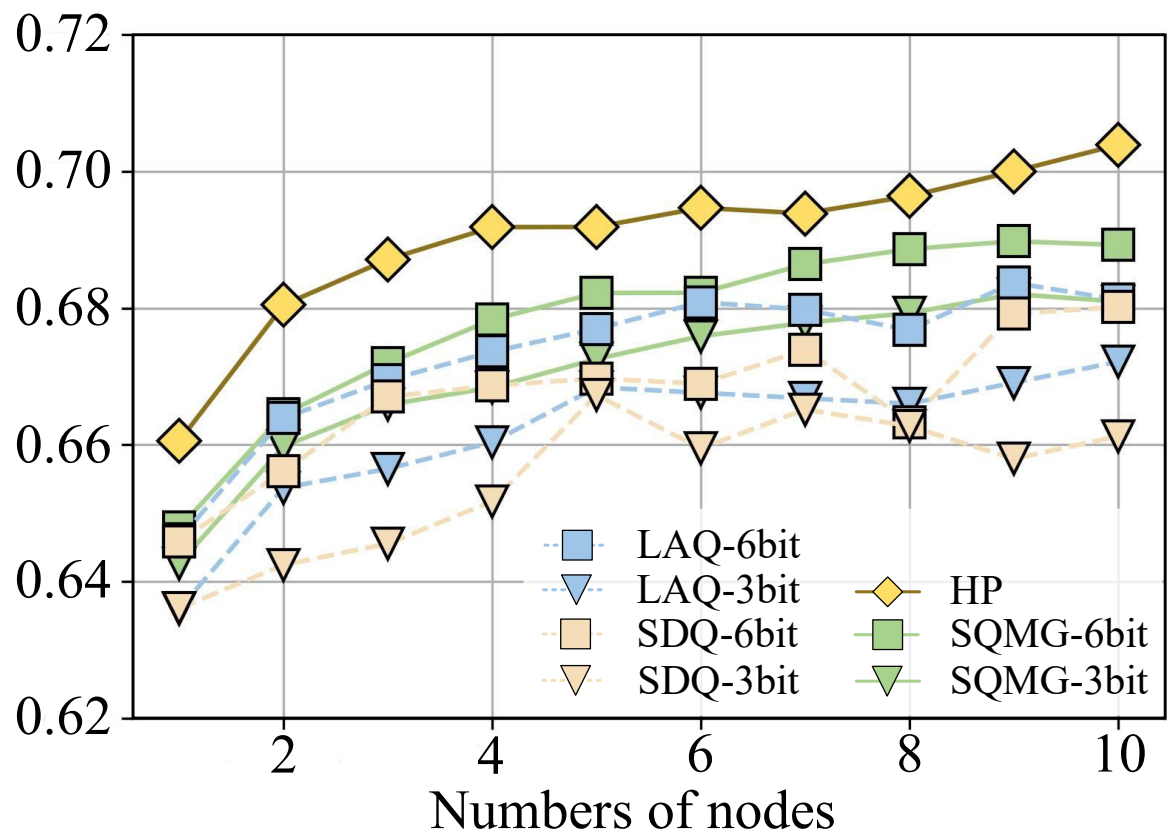
(a) Top-1 Accuracy for local Training Set (First round)



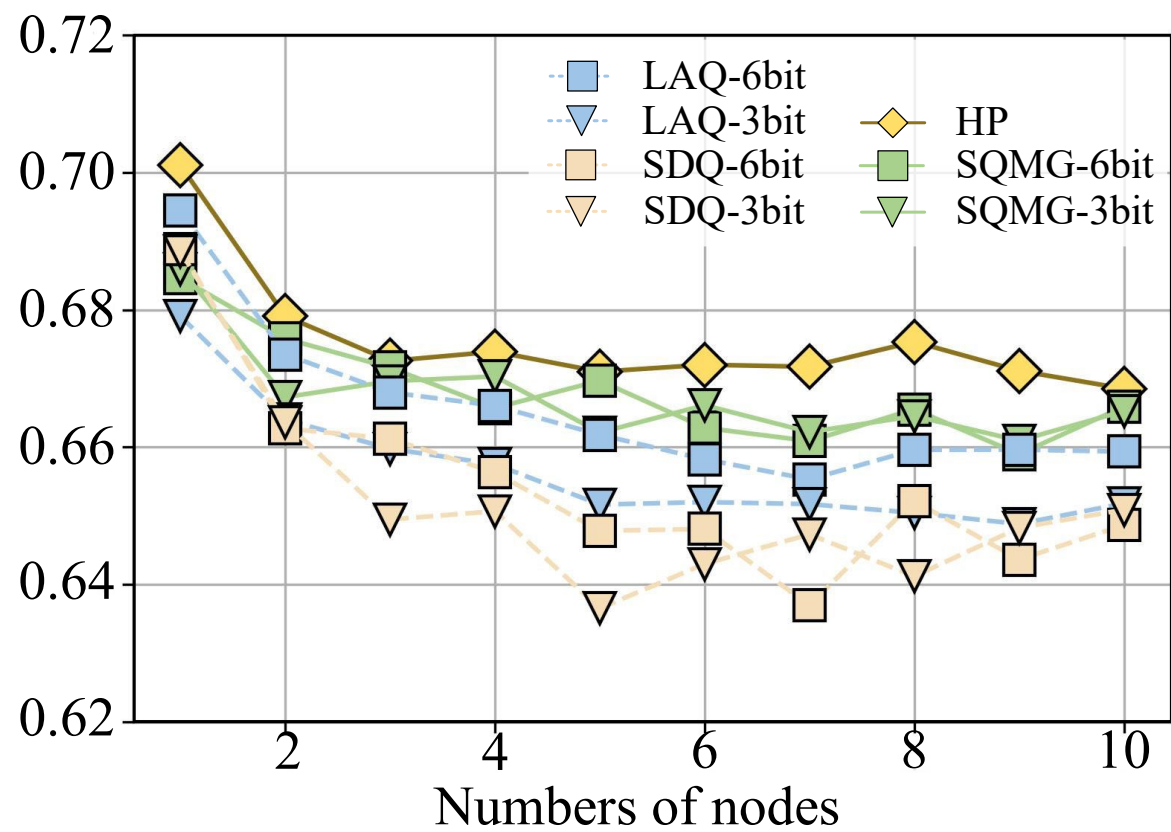
(b) Top-1 Accuracy for Testing Set (First round)

Experimental Results – CNN Model Scenarios

- Top-1 Accuracy of CNN model versus number of nodes after training is completed.
 (a) the local training set; (b) the testing set.



(a) Top-1 Accuracy for local Training Set (Final round)



(b) Top-1 Accuracy for Testing Set (Final round)

Conclusions

- AQSGD utilizes stochastic quantization techniques to reduce communication costs. Conventional quantization methods usually assume a uniform distribution, resulting in the increase of the quantization error.
- This study introduces SQMG approach:
 - SQMG leverages a **multivariate Gaussian model** to capture correlations within the gradient vector, enabling a more comprehensive representation of gradient updates, thus constructing an optimized quantization target space.
 - SQMG introduces a novel **quantization mapping scheme** through an iterative design. The mapping scheme projects the data onto the optimized quantization target space while ensuring that the quantization errors remain below a specified threshold.
- SQMG consistently outperforms existing quantization methods, demonstrating an average of **0.92% and 1.54%** improvements for DNN and CNN training, respectively, while maintaining consistent quantization accuracy across both training and testing datasets with the same bit precision for distributed learning.

IJCNN 2024

June 30-July 5, Yokohama, Japan



Thanks for listening!

Presenter: Hongxing Li

