YAU, Chung Yiu

4th Year Ph.D. student, The Chinese University of Hong Kong
Email: cyyau@se.cuhk.edu.hk Webpage: https://oscaryau525.github.io/
Google Scholar: https://scholar.google.com/citations?user=DXAnvT0AAAAJ

RESEARCH INTERESTS

- ◇ (TMLR) (Arxiv) Data-parallel decentralized optimization for machine learning.
  - – Neural network training on large-scale datasets by utilizing GPU clusters in the setting of fast computation and minimal communication.
- ◇ (ICML 2024) Multi-modal contrastive learning.
- ◇ (ICML 2025) Large language model quantization for memory efficiency and inference speedup.

WORK EXPERIENCE

Applied Scientist Intern **June - Sep 2024**
Amazon Web Services, Santa Clara, California

- ◇ Research on large language model quantization, especially quantization-aware training.
- ◇ Proposed an adaptive rotation method for fine-tuning with paper presented at ICML 2025. Our algorithm fine-tunes a quantized LLM by quantization-aware training on Llama, Pythia and Qwen models.

Applied Scientist Intern **June - August 2023**
Amazon Web Services, Shanghai

- ◇ Study the large batch inefficiency in contrastive learning pre-training.
- ◇ Proposed a small batch sampling algorithm with paper presented at ICML 2024. Our algorithm improves negative pair sampling for pre-training.

EDUCATION

Ph.D. Systems Engineering & Engineering Management **2021 - 2025**
The Chinese University of Hong Kong, Hong Kong. *Supervisor: Prof. Hoi-To Wai*

- ◇ Research focus: analyzing the convergence of novel decentralized optimization algorithms with communication compression.

B.Sc. Computer Science **2017 - 2021**
The Chinese University of Hong Kong, Hong Kong
First Class Honour, ELITE Stream.

ACADEMIC

Invited Talk

Communication Efficient Decentralized Optimization

◇ *Department of ECE, University of Minnesota* **Sep 2024**

LLM Quantization-Aware Training

◇ *INFORMS International 2025, Singapore* **July 2025**

Teaching Assistant **2021 - 2025**

Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong

◇ ENGG2440 Discrete Mathematics for Engineers

◇ FTEC2101 Optimization Methods

Voluntary Reviewer

At conferences and journals:

◇ NeurIPS, ICML, ICLR, TMLR

◇ IEEE TSP, TAC, L-CSS

. . . . . .

RESEARCH PUBLICATION

[1] Chung-Yiu Yau, Haoming Liu, and Hoi-To Wai. A stochastic approximation approach for efficient decentralized optimization on random networks. *arXiv preprint arXiv:2410.18774v2*, 2025.

[2] Quan Wei, Chung-Yiu Yau, Hoi-To Wai, Dongyeop Kang, Youngsuk Park, Mingyi Hong, et al. Roste: An efficient quantization-aware supervised fine-tuning approach for large language models. *Proceedings of the 42st International Conference on Machine Learning*, 2025.

[3] Haoming Liu, Chung-Yiu Yau, and Hoi-To Wai. Decentralized stochastic optimization over unreliable networks via two-timescales updates. *arXiv preprint arXiv:2502.08964*, 2025.

[4] Haoming Liu, Chung-Yiu Yau, and Hoi-To Wai. A two-timescale primal-dual algorithm for decentralized optimization with compression. In *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2025.

[5] Chung-Yiu Yau, Hoi-To Wai, Parameswaran Raman, Soumajyoti Sarkar, and Mingyi Hong. EMC^2: Efficient MCMC negative sampling for contrastive learning with global convergence. In *Proceedings of the 41st International Conference on Machine Learning*, pages 56966–56981. PMLR, 2024.

[6] Chung-Yiu Yau and Hoi-To Wai. Fully stochastic distributed convex optimization on time-varying graph with compression. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 145–150. IEEE, 2023.

[7] Xiaolu Wang, Chung-Yiu Yau, and Hoi-To Wai. Network effects in performative prediction games. In *International Conference on Machine Learning*, pages 36514–36540. PMLR, 2023.

[8] Chung-Yiu Yau and Hoi To Wai. Docom: Compressed decentralized optimization with near-optimal sample complexity. *Transactions on Machine Learning Research*, 2023.

[9] Bingqing Song, Ioannis Tsaknakis, Chung-Yiu Yau, Hoi-To Wai, and Mingyi Hong. Distributed Optimization for Overparameterized Problems: Achieving Optimal Dimension Independent Communication Complexity. *Advances in Neural Information Processing Systems*, 2022.

[10] Qiang Li, Chung-Yiu Yau, and Hoi-To Wai. Multi-agent Performative Prediction with Greedy Deployment and Consensus Seeking Agents. *Advances in Neural Information Processing Systems*, 2022.

[11] Chung-Yiu Yau, Haoli Bai, Irwin King, and Michael R Lyu. DAP-BERT: Differentiable Architecture Pruning of BERT. In *International Conference on Neural Information Processing*, pages 367–378. Springer, 2021.