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

- ◇ 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. (Latest paper: asynchronous algorithm) (TMLR paper).
- ◇ Multi-modal contrastive learning.
  - – Our negative sampling algorithm for pre-training foundation models achieved higher zero-shot accuracy. (ICML paper)
- ◇ Large language model compression.
  - – Fine-tuning a quantized LLM by quantization-aware training with near lossless performance on Pythia and Qwen models. (arXiv paper)
  - – Reduces memory requirement and speedup inference for LLM deployment.

WORK EXPERIENCE

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

- ◇ Research on large language model quantization, especially quantization-aware training in the fine-tuning stage.

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

- ◇ Study the large batch dependence in contrastive learning for uni-modal/multi-modal pre-training.
- ◇ Proposed a small batch sampling algorithm with paper presented at ICML 2024.

EDUCATION

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

- ◇ Analyze the convergence of novel decentralized optimization algorithms with communication compression.

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

ACADEMIC

Visiting Prof. Mingyi Hong's Research Group **Sep 2024**
Department of Electrical and Computer Engineering, University of Minnesota

Teaching Assistant **2021 - Present**
Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong

- ◇ ENGG2440 Discrete Mathematics for Engineers
- ◇ FTEC2101 Optimization Methods

Voluntary Reviewer
ICML 2024 2025, ICLR 2025, TMLR, IEEE TAC, IEEE TSP, IEEE L-CSS

.....

RESEARCH PUBLICATION

[1] 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. *arXiv preprint arXiv:2502.09003*, 2025.

[2] 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.

[3] 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.

[4] Chung-Yiu Yau, Haoming Liu, and Hoi-To Wai. Fully stochastic primal-dual gradient algorithm for non-convex optimization on random graphs. *arXiv preprint arXiv:2410.18774*, 2024.

[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.