

THE UNIVERSITY



OF HONG KONG

Department of Mathematics

Qualifying Research Seminar

Optimization Methods for the Full LLM Lifecycle: Enhancing Efficiency from Pre-training to Deployment

Mr. Xu Zhicheng

MPhil Student, Department of Mathematics, HKU

(Supervisor: Professor Xiaoming YUAN)

March 10, 2026 (Tuesday) at 4:00pm

Rm 210, Run Run Shaw Building, HKU

Abstract

The steadily increasing scale of Large Language Models (LLMs) has posed significant efficiency challenges across their full lifecycle, from large-scale pre-training and task-specific fine-tuning to model deployment. In order to address these challenges, we propose tailored solutions for each stage. First, for pre-training, we introduce PFAdam (Periodically Forgotten Adam) and its decentralized extension, designed to reduce memory footprint and communication overhead in distributed settings. Second, for parameter-efficient fine-tuning, we propose RA-SpaRC (Robust Adaptation with Sparse Plus Low-Rank Compressors), a novel initialization method that achieves high resource utilization efficiency by dynamically allocating the parameter budget between sparse and low-rank components. Third, we extend the core principle of RA-SpaRC to model pruning by introducing SpaRk-Pruner (Sparse Plus Low-Rank Matrix Decomposition Pruner), a novel method for semi-structured compression that enhances inference efficiency. This comprehensive approach aims to reduce the computational resources required for developing and deploying LLMs while maintaining competitive performance.

All are welcome