For favour of Posting

THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Department of Mathematics / Dept of Industrial Engineering & Decision Analytics

JOINT SEMINAR ON STATISTICS AND DATA SCIENCE

Polar Gradient Methods: A Unifying Preconditioning Perspective on Matrix-Gradient Methods for Deep Learning Optimization

By

Dr. Tim Tsz-Kit Lau

DRW and University of Pennsylvania

Abstract

The ever-growing scale of deep learning models and training data underscores the critical importance of efficient optimization methods. While preconditioned gradient methods such as Adam and AdamW are the de facto optimizers for training neural networks and large language models, structure-aware preconditioned optimizers like Shampoo and Muon, which utilize the matrix structure of gradients, have demonstrated promising evidence of faster convergence. In this work, we introduce a unifying framework for analyzing "matrix-aware" preconditioned methods, which not only sheds light on the effectiveness of Muon and related optimizers but also leads to a class of new structure-aware preconditioned methods. A key contribution of this framework is its precise distinction between preconditioning strategies that treat neural network weights as vectors (addressing curvature anisotropy) versus those that consider their matrix structure (addressing gradient anisotropy). This perspective provides new insights into several empirical phenomena in language model pre-training, including Adam's training instabilities, Muon's accelerated convergence, and the necessity of learning rate warmup for Adam. Building upon this framework, we introduce polar gradient methods (PolarGrad), a new class of preconditioned optimization methods based on the polar decomposition of matrix-valued gradients. As a special instance, PolarGrad includes Muon with updates scaled by the nuclear norm of the gradients. We provide numerical implementations of these methods, leveraging efficient numerical polar decomposition algorithms for enhanced convergence. Our extensive evaluations across diverse matrix optimization problems and language model pre-training tasks demonstrate that PolarGrad outperforms both Adam and Muon. This talk is based on joint work during the speaker's postdoc with Prof. Qi Long and Prof. Weijie Su from University of Pennsylvania.

Bio: Dr. Tim Tsz-Kit Lau is an AI Researcher at DRW Palo Alto. Prior to his role at DRW, he was a postdoctoral researcher at the University of Pennsylvania from November 2024 to May 2025, and at the University of Chicago Booth School of Business from October 2023 to October 2024. He received his Ph.D. in Statistics from the Department of Statistics and Data Science, Northwestern University in 2023. In his published research, he has worked broadly at the interface of statistics, machine learning and optimization, with applications to large language models (LLMs) and more broadly generative AI (GenAI). His current research aims to advance state-of-the-art optimizers and training strategies for scaling and stabilizing language model training in a theoretically principled way, improving both pre-training and post-training scaling and training stability.

Date : 20 January 2026 (Tuesday)
Time : 10:30a.m.-12:00noon
Venue : Room 4472 (near Lift 25/26)

All are welcome