For favour of Posting

THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Department of Mathematics / Dept of Industrial Engineering & Decision Analytics

JOINT SEMINAR ON DATA SCIENCE AND APPLIED MATHEMATICS/STATISTICS

Symmetry-Compatible Principle for Optimizer Design: Embeddings, LM Heads, SwiGLU MLPs, and MoE Routers

By

Dr. Tim Tsz-Kit LAU
DRW and University of Pennsylvania

Abstract

A striking geometric disparity has long persisted in the practice of deep learning. While modern neural network architectures naturally exhibit rich symmetry and equivariance properties, popular optimization methods, such as Adam and its variants, operate inherently coordinate-wise, rendering them unable to respect the equivariance structures of the parameter space. In this paper, we address this disparity by introducing a symmetry-compatible principle for optimizer design. Specifically, we argue that the gradient update rule should be equivariant under the symmetry group acting on the corresponding weight block, and should remove symmetry-redundant directions when the parameterization has quotient structure. Following this principle, we first provide a unified perspective on bi-orthogonally equivariant updates for general matrix layers, as employed by stochastic spectral descent, Muon, Scion, and polar gradient methods. More importantly, by moving from orthogonal groups to permutation and shared-shift symmetries, we derive new classes of symmetry-compatible optimizers tailored to parameter blocks whose symmetries differ from those of ordinary matrix layers: for embedding and LM head matrices, left-permutation and right-orthogonal equivariance leads to one-sided spectral, row-norm, and hybrid row-norm/spectral updates, with projected variants for LM heads; for SwiGLU MLP projections, intermediate-neuron permutation symmetry motivates row-aware and column-aware variants; and for MoE routers, expert-permutation symmetry together with shared-logit-shift invariance gives rise to projected centered row-norm and left-spectral updates. These constructions yield an end-to-end layerwise optimizer stack in which each major matrix-valued parameter class is assigned an update whose equivariance matches its symmetry group. We corroborate this optimizer design principle through extensive pre-training experiments on dense and sparse MoE language models, including Qwen3-0.6B-style, Gemma 3 1B-style, OLMoE-1B-7B-style, and downsized gpt-oss architectures. Across these experiments, symmetry-compatible update rules consistently improve final validation loss, reduce expert load imbalance in sparse MoE models, and in several cases control final vocabulary-logit growth, improve router stability, and overall training stability over the corresponding AdamW updates. This talk is based on the joint work arXiv:2605.18106 with Prof. Weijie Su.

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 and Data Science from Northwestern University in 2023. He also received an MSc in Big Data Technology from HKUST in 2017, and a BSc in Mathematics and a Master of Statistics from HKU in 2015 and 2016 respectively. Dr. Lau 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 focuses on developing theoretically principled optimizers and training strategies for scaling and stabilizing language model training, with the goal of improving both pre-training and post-training efficiency, scalability, and training stability.

Date : 29 June 2026 (Monday)
Time : 3:00p.m. – 4 :00p.m.
Venue : Room 2302 (near Lift 17/18)

All are welcome