

KANG AN

Email: ka56@rice.edu; [LinkedIn](#); [Google Scholar](#); [GitHub](#)

EDUCATION

Rice University

PhD in Computational Applied Mathematics and Operations Research

Houston, TX

Sept 2022 – Present

New York University

Master of Science in Mathematics, Courant Institute of Mathematical Sciences

New York, NY

Sept 2020 – Jun 2022

Sichuan University

Bachelor of Science in Applied Mathematics, Wu Yuzhang Honor College

Chengdu, CN

Sept 2015 – Jul 2020

INDUSTRIAL EXPERIENCE

Meta, Research Scientist Intern, Co-Design Team

May 2026 - Aug 2026

- Identified a **dimension-dependent update-scale mismatch** between KL-Shampoo and RMSProp; evaluated **Update Normalization** and **Preconditioner Normalization** as alternatives to learning rate grafting, matching or slightly better performance in the tested LLM setting.
- Investigated **Muon for QKV parameters** by comparing head-count and head-dimension scaling in LLaMA3 models. After learning-rate tuning, Head-Split Muon **retained validation-loss gains over tuned AdamW** across the tested per-head matrix aspect ratios, while optimal learning rates differed between the two scaling strategies.
- Formulated **time-averaged semi-orthogonality in expectation (TAOE)** to study adaptive update geometry. Evaluated it in full-batch ViT training on TinyImageNet and, separately, in stochastic LLaMA3 training using TorchTitan, finding that **TAOE anisotropy alone** did not reliably predict optimizer performance or identify optimal β_2 values.

RESEARCH EXPERIENCE

Adaptive Structured Gradient Optimization in Deep Learning

Dec 2024 – Jul 2025

- Proposed **ASGO** and established **improved convergence rates** over Shampoo under the analyzed assumptions, while reducing **optimizer-state memory usage by over 50%** relative to Shampoo.
- Implemented a practical ASGO variant with **8% wall-clock overhead** relative to AdamW; characterized its relationship to Muon through an analogy with Adam and SignSGD.
- Evaluated ASGO on **GPT-2 pre-training**, achieving lower **validation loss** than AdamW and Shampoo in the evaluated setting.
- The paper ASGO: Adaptive Structured Gradient Optimization has been accepted to **NeurIPS 2025**.

Demystifying Manifold Constraints in LLM Pre-training

Jan 2026 – Aug 2026

- Developed **MACRO** and a constraint-radius selection principle; identified functional overlap between manifold constraints and **RMSNorm**. Observed validation loss to be largely insensitive to changes in constraint radius over the tested range, with learned normalization scales adapting accordingly.
- Showed that manifold constraints directly regulate **weight-rotation dynamics**. It fulfills the role of traditional weight decay and offers a principled alternative.
- Demonstrated substantially **less performance degradation than Muon** when switching from FP32 to **BF16 master weights** without stochastic rounding, and provided an explanation based on a **Gaussian coordinate-threshold model**.
- Established convergence guarantees for MACRO and evaluated it on **Qwen3-based models (240M–1.2B parameters)** using TorchTitan, demonstrating consistent gains over the tested unconstrained baselines and performance on par with SOTA optimizers.

Basis Adaptation Training Through Curvature-Informed Descent

May 2026 – Present

- Proposed a basis-adaptation optimizers framework and derived **BATH**, with theoretical analysis of its potential advantages over existing rotation-based optimizers.
- Implemented **BATH** in TorchTitan with **ZeRO-1 optimizer-state sharding** and evaluated BATH and KL-BATH on **1B Llama 3 dense models**, demonstrating consistent gains over Muon and SOAP in the evaluated settings.

TECHNICAL SKILLS

- **Languages & Tools:** Python, MATLAB, Linux, Git, $\text{\LaTeX}$
- **Frameworks & Distributed Training:** PyTorch (DDP/FSDP), TorchTitan, DeepSpeed.
- **Research skill:** LLM pre-training, Optimization, Deep-Learning model training.